

Fighting malaria from the front

Too much of malaria research is piecemeal, and the organizations that are supposed to support it are insufficiently effective. Better focusing could yield more funds.

"Fosters a 'not my problem' culture;" "Repeatedly fails to deliver on targets;" "Blames others;" "Makes excuses." These unfavourable indicators typical of modern performance management ring too true in relation to the organizations that are fighting malaria. Although much has been achieved, international and national bodies, not least the African governments concerned, need to be held more accountable for their actions—or inaction. More rigorous and frequent evaluation and oversight at all levels is needed.

Resources are scarce, making it doubly essential to define priorities and funding in the war against this devastating disease. The troops are willing, as testified by the gathering last week in Arusha, Tanzania, of some 950 of them (see page 351), including over 400 Africans. There were regiments representing the myriad skills required to secure a victory: basic researchers who are developing tomorrow's weapons, and helping today's to be used more effectively; special forces and infantry, charged with the groundwork in control and the clinic; and strategists and paymasters, who ultimately control many of the purse and political strings.

Yet good leadership is too often lacking. A recent independent review slammed the international community's flagship control effort, Roll Back Malaria (RBM), which was launched in 1998 by the World Health Organization (WHO) with help from country donors, the World Bank and several arms of the United Nations. RBM helped to bring malaria from the political wilderness to the international spotlight, but has failed to deliver adequate progress in controlling the disease (see *Nature* **419**, 422; 2002).

Responsibility for RBM is being transferred from the bureaucratic WHO to an independent steering group. Officials from RBM, the World Bank and others privately blame the lack of progress on each other. The new full-time executive director of RBM should be an exceptional individual capable of calming tensions, focusing the job and getting things done. The fact that widespread scepticism about RBM's effectiveness was not officially acknowledged sooner raises concerns about the degree of oversight in international health.

A brighter picture has emerged from the first five-year review of the Multilateral Initiative on Malaria (MIM), an international consortium of research agencies and funders that aims to beef up malaria research and cooperation in Africa. The MIM has notched up some successes, including nurturing several emerging pan-African research networks. But the review also recommends "strongly" that the MIM focus more on defining clearer overall objectives and target 'big picture' business plans, which may be better able to attract significant funding.

What applies to the MIM also applies to the malaria research community in general. Investigator-driven research is important, but the scarcity of funds means that it risks being piecemeal. Scientists' efforts to seek more funds by collaborating to create larger, well-defined projects with clear milestones and goals should be encouraged. Research administrators also need to be imaginative in creating flexible funding arrangements that go beyond traditional institutional boundaries. The palpable buzz in Arusha was a good sign. As the war on malaria moves forwards, especially in Africa, where most deaths occur, the troops and their generals must embrace a more professional approach. ■

Postdoctoral abuse (cont.)

The failure to provide appropriate career structures for young scientists persists.

In these enlightened times, one might expect an investigation that revealed the miserable existence of thousands of workers, who are exploited and forced to work repeatedly on contracts for low pay with few benefits or little job security, to raise at least ripple of protest. But when a British parliamentary committee highlighted the continuing plight of postdoctoral researchers last week, and pointed out that university research is now second only to the fast-food industry in the proportion of casual labour it employs, a pricked conscience or two seems to have been the best that it could hope for.

The plight of the postdoc has been well documented, not least in the pages of this journal. In recent years, *Nature* has covered a depressing range of reports and studies from Britain as well as France, Germany, the United States and others, all saying that postdocs are undervalued and poorly paid. There have been studies, initiatives and pilot schemes aimed at improving the situation. Why, then, does so little seem to have changed?

Often compared to an oil tanker altering course because of its slow pace, change in higher education often more closely resembles the geological formation of the oil itself. With little incentive to update employment practices that discriminate against those on fixed-term contracts, there seems to be little chance of achieving the radical

change in university culture demanded by British parliamentarians (see <http://www.parliament.uk/commons/selcom/s&thome.htm>).

Yet with few guaranteed sources of even medium-term income, the universities will rightly argue that they have at least one hand tied behind their back. Those providing the funds must also shoulder more responsibility for the well-being of those who do the research. The British government's recent decision to raise postdoc salaries by £4,000 (US\$6,250) a year is welcome. All research councils should make it possible for contract researchers — many of whom have years of experience — to apply for grants in their own name.

The well-documented poor working conditions have not given rise to an international recruitment or retention crisis. A famously stoical bunch, many researchers only look as far as the next research position. What some still view as a rite of passage before a glittering career has effectively become a career in itself.

Generally, established researchers have failed to confront this issue. Too many still hold the out-of-date opinion that postdoctoral years are a 'sink or swim' exercise that "never did me any harm". On the contrary, there needs to be better management of postdoctoral careers, and better mentoring. Funding agencies should take such demonstrable commitments, or a lack of them, into account. ■





Closing the gates
US visa rules tighten controls on visiting researchers
p349



Cold shoulder
Colleges drop plans to merge in face of stiff opposition
p350



Focus on facts
Malaria conference homes in on poorest countries
p351



Animal passion
University fights to set up primate-research centre
p353

Ecological riches threatened as oil-spill history repeats itself

Monica Salomone, Madrid

"It is as if we have not learned anything from the past," says Ramon Núñez, director of the Aquarium Finisterrae in La Coruña on Spain's northwestern coast. Núñez's aquarium, home to 35,000 fish and countless marine invertebrates, was established three years ago in response to an oil spill that hit the surrounding coastline in 1992. It is now the symbol of a new disaster.

The animals it houses cannot survive without the water flowing in through the sluice gates that connect the aquarium to the sea. But heavy fuel oil, washed ashore from the wreck of the tanker *Prestige*, is flowing through the gates and threatening the fish inside.

Núñez is waiting to see how much more oil will escape from the tanker, which sank 250 kilometres offshore on 19 November. Between 10,000 and 20,000 tonnes had found its way into the open ocean, contaminating 400 kilometres of coastline, as *Nature* went to press. Further slicks are expected to be washed ashore this week. "We are yet to see the worst of it," says Victoriano Ugorri, a marine zoologist at the University of Santiago de Compostela.

Some 60,000 tonnes of oil went down with the *Prestige*, which broke in two before sinking. Oceanographers estimate that the temperature on the sea floor, 3.6 kilometres below the surface, is 2–3 °C. The heavy oil on board the tanker freezes at 6 °C and so is unlikely to leak out. But planes flying over the site of the wreck last weekend reported that large new slicks have appeared where the vessel sank. The French government has sent the *Nautille*, a deep-water research submarine, to assess the situation.

The Spanish Society of Ornithology says that several hundred birds have already been treated for oil contamination. Almost 20 bird species from the Atlantic and the Mediterranean feed in the area at this time of year. Conservationists are particularly worried about the Iberian guillemot (*Uria aalge ibericus*), of which fewer than 20 pairs are known to exist in the area.

Environmental organization the World



Beached: northwestern Spain is struggling to cope with the effects of oil from the wrecked *Prestige*.

Wide Fund for Nature (WWF) is also concerned about the spill's potential effect on the Galician Bank, an ecologically rich zone located 200 km off the Spanish coast. Only a few months ago, the WWF asked for the area to be given special protection.

Ugorri estimates that the region will take at least five years to recover, even if the extra oil remains on the sea floor. He and colleagues have studied the impact of the 1992 spill, which came from the *Aegean Sea* tanker. "Most of the species affected at the time have more or less recovered, but they needed a decade," he says.

This spill has spread over a much larger area, although the covering of oil is thinner, which could speed up recovery. "The less affected areas will act as reservoirs for species," says Ricardo Prego, head of the marine biogeochemistry group at the Institute for Marine Studies in Vigo.

The Spanish government is using mechanical cleaning methods, such as

removing oil-covered sand from beaches and using high-pressure water jets to clean rocks. But the fuel contains over twice the normal level of sulphur, which is toxic and makes the waste difficult to handle. It is also very dense, and so evaporates more slowly than oil from other spills.

Debate over the government's handling of the disaster is likely to last at least as long as the clean-up operation. Spain's northwestern coast has suffered six major spills in the past three decades. Yet environmental organizations and marine scientists say that the authorities do not have enough cleaning equipment and ships to tackle the latest spill. Critics also allege that the government has deliberately underestimated the amount of oil that has been washed ashore.

As the argument rumbles on, Núñez is keeping his aquarium open. "It is the only place where we can show to the world the dying ecosystem," he says. ■

SANTIAGO LYON/AP

Japan ponders steps to probe data errors

David Cyranoski, Tokyo

Where does the border lie between simple error and scientific misconduct? That is the question thrust in the face of the Japanese research community by the unsavoury means of an anonymous letter.

The letter was sent to funding agencies and publishers throughout Japan, Europe and the United States in August 2001. It concerned papers by Takuji Tanaka, a cancer researcher at Kanazawa Medical University. The anonymous author claimed that there were some unusual similarities between figures in an article on foods that may help to prevent cancer published in September 1997 in the *Japanese Journal of Cancer Research* (T. Tanaka *et al.* *Jpn J. Cancer Res.* 88, 821–830; 1997) and those in a 1999 paper in *Carcinogenesis* (T. Tanaka *et al.* *Carcinogenesis* 20, 1477–1484; 1999). The letter says that both papers include two identical figures for polyamine compounds in rats, even though the two experiments used different numbers of rats for different lengths of time. The letter went on to allege that this was just one of at least 30 discrepancies in articles by Tanaka.

In January 2002, the *Japanese Journal of Cancer Research* ran a correction to the 1997 paper (*Jpn J. Cancer Res.* 93, 109; 2002). “It was a clear mistake,” admits Tanaka, “but I have never fabricated data.” He also denies any misconduct in the other cases.

Tanaka's supporters say that he is the target of an unjustified personal attack, and

they reject the anonymous author's claim that he or she is acting on behalf of an unnamed watchdog group dedicated to protecting scientific integrity. The author's allegations are limited to Tanaka. “He or she just wants Tanaka to lose his job,” says Hideki Mori, head of the medical department at Gifu University and a co-author on the two papers mentioned in the letter. “Tanaka's reputation has suffered unfairly.”

Mori suggests that Tanaka should take the case to a lawyer or the police. “The author of the letter is like a stalker,” notes Mori, who claims to have a very good idea of the author's identity.

But it would be healthy for the scientific community to have a mechanism in place to investigate such charges properly, Mori points out. Japan lacks anything like the US health department's Office of Research Integrity, and officials at both the ministry of education and ministry of health say that there is no system in place to look at cases of misconduct.

Masaaki Terada, editor of the *Japanese Journal of Cancer Research* and a former director of the National Cancer Center in Tokyo, says that he did wonder whether the allegations should be raised with anyone other than his journal's executive editors. He agrees that Japan needs a system to investigate misconduct accusations, and says that he plans to raise the question with the health ministry. “There should be a government-

level body to deal with this,” says Terada.

The problem of misconduct is especially difficult in Japan because the society does not look kindly on insiders who blow the whistle, says Tokai University's Kiyoshi Kurokawa. “Even if you want to report a problem to your professor, he may be afraid of his dean or hospital director, who will be afraid of the education ministry,” says Kurokawa. “Such reporting may backfire.”

Some critics say that the reluctance to criticize others, especially those in established positions, was to blame for one of Japan's most notorious misconduct case — the 25 years of research in which an archaeologist pushed back Japan's stone age some 600,000 years with his ‘discoveries’ of planted relics. After the fraud came to light, many researchers claimed that they had doubted the findings all along.

More broadly, a spate of scandals in industry and government circles has brought the problem of encouraging and protecting whistleblowers to centre stage. The cabinet office is currently considering a system to protect informants.

Kurokawa, who is vice-president of the Science Council of Japan, an organization that represents thousands of scientists, has established a committee to come up with guidelines on the subject of scientific misconduct. Such a system could also offer protection from false accusations, Mori points out. ■

Additional reporting by Oliver Schmidt.

Maths adds up for Berlin centre

Quirin Schiermeier, Munich

A centre that will apply mathematics to problems in physics, biology, telecommunications and transportation has opened its doors in Berlin. Advocates say that it could catapult the German capital to the top of the league in mathematics research.

Over the next four years, the DFG Research Centre for Mathematics for Key Technologies will receive 5 million euros (US\$5 million) per year from the DFG, Germany's main research funding agency, and a further 3 million euros per year from Berlin's state government.

“This is big money for mathematics,” says Martin Grötschel, vice-president of the Konrad Zuse Centre for Information Technology in Berlin and spokesman for the new centre.

“The nature of our discipline allows us to invest the bulk of the money in people, rather than in equipment,” says Grötschel. Six new full professors, seven young

principal investigators and 60 other scientific staff will be hired at the centre in the coming months.

Günter Ziegler, a mathematician at the Technical University of Berlin, who last year won the DFG's prestigious Leibniz prize, is enthusiastic about the new investment. He says that the funding boost should help to block the brain drain of mathematicians from German academic positions to industry and to the United States, where mathematics is winning increased support from the National Science Foundation (see *Nature* 414, 676; 2001).

Any concern that the new DFG research centre's emphasis on applications will be at odds with fundamental research is groundless, says Grötschel. “Modelling common problems in industry, banking and science will require new algorithms,” he says. “This will trigger a loop of research that is equally beneficial” to both theoretical and practical mathematics. ■

Summing up: computer modelling will help to boost biology at Berlin's maths institute.

Visa clampdown hits home at US universities

Kendall Powell, Washington

Tighter immigration rules are causing a broad increase in visa delays and refusals for students and postdoctoral fellows seeking entry to the United States, two surveys show. And the problems are set to get worse in the new year, university officials say, when a scheme for tracking the students' movements is due to come into force.

"There is an intrinsic tension between advancing scientific knowledge and national security," says Irving Lerch, director of international affairs at the American Physical Society (APS). "Right now, the argument has come down on the side of security, but in the long run that can be deleterious to our research system, economy and security development."

Although official figures indicate that there was no drop in overall international student enrolment in the United States in the immediate wake of the terrorist attacks on 11 September 2001, the two surveys, conducted separately by the Association of American Universities (AAU) and the APS, indicate a clear effect on those arriving for the 2002–03 academic year.

The number of visiting researchers on 'J

visas', for exchange visitors, dropped by 11% on average at 20 major US universities, the AAU finds. And 20% of international physics students were denied a visa after being accepted by the 80 departments questioned by the APS.

The surveys indicate that students and postdocs from China, Russia, India, Pakistan and the Middle East were affected most, as immigration officials made use of existing powers to deny visas, for example, if they have reason to suspect that a visa applicant intends to stay in the country permanently.

As a result, researchers say, projects have been postponed, teaching-assistant positions have gone unfilled, some students have gone elsewhere, and scholars have missed scientific conferences. The effect on US universities has been substantial: in 2001, foreign students gained almost half of all US engineering and physical-sciences doctorates and a quarter of those awarded in the life sciences.

But the biggest worry for universities is a looming deadline for the introduction by the Immigration and Naturalization Service of the computerized Student and Exchange Visitor Information System (SEVIS). The system, which has been in development since a stu-



Customary wait: heightened security is hampering foreign students' entry to the United States.

dent was suspected of involvement in the 1993 World Trade Center bombing, aims to track the whereabouts of several hundred thousand holders of F (student), J and M (vocational) visas. It will eventually connect educational institutes, consular offices and ports.

Terry Hartle of the American Council on Education, which represents education interests in Washington, says that implementing this system is a "moonshot" for the immigration service, and claims that good progress has been made on it. But some observers say the system will struggle to cope — and may well crash altogether — as institutions rush to join before the deadline on 31 January 2003.

Anxious university admissions officers say they haven't received full details of SEVIS' regulations. "It's going to be down to the wire," says John Pearson, director of the Bechtel International Center at Stanford University in California. He and others worry not just about breakdowns of the system leading to further delays for students, but about serious repercussions for universities that fail to meet SEVIS' requirements. Under current rules, such institutions could lose all international students, and proposed legislation could also strip them of federal research funding.

Muslim students who already have visas, meanwhile, are complaining of harassment upon their return to the United States from vacations. "This is having an impact on even the best science students' ability to focus on their work," says Altaf Husain, president of the Muslim Student Association of the US and Canada. "They are here working diligently, usually as a model student and citizen on campus, and then have to go through the humiliating experience" of an interrogation.

Back in May, John Marburger, President Bush's science adviser, proposed the creation of an Interagency Panel on Advanced Science Security to help to review visa applications. But according to Lerch, the proposal is in limbo pending the organization of the new Department of Homeland Security.

M. SEROT/REUTERS

Britain failing to bar risky students

Natasha McDowell, London

Rihab Taha, who headed Iraq's biological weapons programme, is one product of the UK university system that the British government would sooner forget. But according to an investigation by the BBC, a scheme designed to keep similar individuals out of British universities may be failing.

The Voluntary Vetting Scheme was set up in 1994 after it was revealed that Taha spent five years as a postgraduate microbiologist at the University of East Anglia in Norwich before taking up her post in Iraq.

The 95 universities and centres of higher

education that have signed up to the scheme take government advice on whether to accept students for certain courses, ranging from biology to computer science, from countries such as Iraq, Iran and Pakistan. The government has provided guidance on 600 students this year, and has asked institutions to refuse 24 applicants.

But an investigation by the Radio 4 current affairs programme *File on 4*, details of which were broadcast last week, revealed that many universities do not cooperate with the scheme. Of the 41 institutions with microbiology departments that replied to *File on 4*'s survey, 12 said that they did not comply. Their reasons included too much paperwork and a lack of confidence in the system. Four universities said that they had never heard of the scheme.

The UK Foreign and Commonwealth Office, which runs the vetting scheme, says that it works well for a voluntary system. "We are not claiming that all universities cooperate, but 100% of those we consider of greatest concern have signed up," says a spokesperson. A compulsory scheme — an option favoured by some members of parliament on the Foreign Affairs Select Committee — would be unlikely to catch many more people and would cost a lot more to administer, the spokesperson adds.



Manchester merger set to proceed as southerners go solo

David Adam, London

To merge or not to merge? That is the question on which British universities can't appear to make up their minds.

While the heads of Imperial College and University College London (UCL) were dodging the slings and arrows of outraged staff, those at the University of Manchester and the University of Manchester Institute of Science and Technology (UMIST) have calmed their sea of troubles sufficiently to start ordering new stationery.

The governing councils of both Manchester and UMIST have now given their blessing to a merger that could be completed as soon as autumn 2004 (see *Nature* 416, 114; 2002). Officials at both argue that the move will create "a truly world-class" institution in Manchester.

But for UCL and Imperial, the nuptials are well and truly cancelled. In statements issued on 18 November, the colleges said that following "an intense period of deliberation" they concluded that "the best interests of the two institutions are not served by a formal merger".

The London colleges first announced in October that they were considering joining forces to form a behemoth with twice the research income of Oxford and Cambridge universities (see *Nature* 419, 658, 2002; and Correspondence, page 359 of this issue). The colleges' heads, Richard Sykes at Imperial and Derek Roberts at UCL, argued that a merged institution would also be more globally competitive.

But critics were quick to pour scorn on the plans, which many staff and students at UCL viewed more as a takeover than a union of equals. UCL pharmacologist David Colquhoun calls the proposal's collapse a victory for "e-democracy" — hundreds of researchers signed online petitions opposing the merger, and both staff and students set up protest websites.

The student website even claimed the posthumous support of UCL founder



Jeremy Bentham: a dim view from the grave.

Jeremy Bentham, who died in 1832 and whose preserved remains are displayed at the college. It quotes him thus: "It was ever the case in this world, alas, that what has been painstakingly constructed over centuries may be reduced to rubble within moments."

Venter aims for maximum impact with minimal genome

Erika Check, Washington

Not for the first time, geneticist Craig Venter's latest wheeze has set US biology abuzz. This time he has reignited the debate over open publication of research results by declaring that he may not release all of the details of his new project.

On 21 November, Venter said he intends to synthesize a bacterial genome from scratch, and then insert it into a cell to see if it can direct the normal functions of the organism. Some observers claim that this process will effectively create a new life-form, but geneticists argue that Venter's project stops far short of this, because he will insert the genome into a naturally occurring, albeit modified, cell.

Venter also says that he may not publish the methods for this work, in case they were used to make biological weapons. "Depending on what is happening in this field, we may not disclose all these details," he says.

The planned work will build on an earlier project in bacterial genomics that he started at The Institute for Genomic Research in Rockville, Maryland. In that project, a team led by molecular geneticist Clyde Hutchison knocked out genes from a *Mycoplasma genitalium* bacterium one by one, and then estimated how many of the bacterium's 517 genes are required for it to live. The researchers estimated that about the bacterium needs about 300 genes to survive (C. A. Hutchison *et al. Science* 286, 2165–2169; 1999).

Venter's new project will also focus on what constitutes a 'minimal genome'. His approach will be different, however: he intends to construct a minimal *Mycoplasma* genome by building its chromosome from scratch. Scientists at Venter's Institute for Biological Energy Alternatives will use chemical synthesizers to manufacture the genes that should be required for *Mycoplasma* to live. Earlier this year, Eckard Wimmer of the State University of New York at Stony Brook used a similar technique to create an infectious poliovirus, which is much simpler than a bacterium (J. Cello, A. V. Paul and E. Wimmer, *Science* 297, 1016–1018; 2002). But Venter plans to take this process a step further, by putting the genes into a *Mycoplasma* cell with the original genetic material removed, to see if the new chromosome works.

If all goes according to plan, Venter will attempt to add further genes to turn the minimal *Mycoplasma* into something useful — perhaps a fuel or an agent for environmental remediation. The US Department of Energy has given his alternative energy institute \$3 million over three years to work towards this goal, and Venter has hired Nobel laureate Hamilton Smith, who worked with him at



God news? Critics of Craig Venter's project claim that he is planning to create a new form of life.

Celera Genomics, to lead the project.

The proposed task will be difficult, but not impossible, scientists commented last week. They said that Venter's main challenges will be synthesizing long, accurate stretches of DNA, and putting the finished synthetic chromosome into a cell.

Most dismissed claims that Venter is playing God by trying to create a new kind of organism. "Geneticists have been trying to get existing life forms to do things they wouldn't ordinarily do for a long time, and it's called genetic engineering," said George Weinstock, co-director of the Human Genome Sequencing Center at Baylor College of Medicine in Houston, Texas.

But some cried foul at statements made by Venter to the effect that he might not publish full details of his methods. Earlier this year, he strongly criticized Wimmer's group for publishing the details of its work (see *Nature* 418, 265; 2002). Venter stands by his criticism of that project, but says that "the community needs to be cautious" about what it publishes.

Venter's critics argue that he should not restrict publication without consulting the wider scientific community. "It's not for him to say," says Fred Blattner, a geneticist at the University of Wisconsin-Madison who is constructing minimal genomes in the bacterium *Escherichia coli*. "He ought to fall on the side of releasing the data unless a consensus develops that he shouldn't do it."

Malaria initiative cries out for action in Africa

Declan Butler, Arusha, Tanzania

The fight against malaria must shift its focus to the poorer African nations that suffer most from it, participants at the largest-ever international meeting on the disease were told last week.

Together, these countries account for 90% of the world's malaria deaths, attendees at the third pan-African conference of the Multilateral Initiative on Malaria (MIM) in Arusha, Tanzania, heard.

"Africa itself needs to take the lead in the challenge," says Ebrahim Samba, regional director for Africa for the World Health Organization (WHO). "If we are asking donor countries to write off our debts, we should commit ourselves to using some of the money to combat malaria. We should also use money that is now being spent on arms."

The meeting brought together 950 experts in all aspects of malaria research and prevention, from physicians, cell biologists, geneticists and vaccine developers to economists, policy-makers and financiers.

Almost half of the participants came from 29 African countries. Such a broad gathering would have been unlikely only a few years ago, when malaria research was languishing after decades of political and scientific neglect. But in 1997, a small group of research agencies, charities, aid donors and scientists set up the MIM and held their first conference in Dakar, Senegal, to explore ways forward and to plan a multifaceted assault on the disease.

Since then, action against malaria has acquired a higher profile, and several international initiatives have been created for malaria research and control, and for drug and vaccine development. The Arusha meeting took stock of their progress — as well as that of the pledges made by African heads of state at a meeting in Abuja, Nigeria, in April 2000. They had pledged to take decisive steps towards halving the world's malaria burden by 2010, and to ensure that 60% of those affected have access to treatment, protection during pregnancy, and are protected by insecticide-treated bednets.

These promises were made when the African leaders signed up to Roll Back Malaria (RBM), a global initiative created in 1998 by Gro Harlem Brundtland, director-general of the WHO. RBM is now being revamped in line with the recommendations of an external review (see *Nature* **419**, 422; 2002). It will no longer be controlled by the WHO, and its secretariat will report instead to a governing board representing funders of the effort and the countries wracked by the disease. It will focus its efforts on a small number of malaria-plagued countries where its ideas stand a decent chance of being put into practice quickly.

For, despite the RBM, malaria is still killing



Waiting for help: across the poorer African nations, thousands die from malaria every day.

thousands of children every day. "We need to go faster," says Gerald Keusch, director of the Fogarty International Center at the US National Institutes of Health in Bethesda, Maryland, and head of the MIM secretariat. With no vaccine in sight for at least ten years, Keusch shares the view that the most pressing need is to have adequate control measures on the ground. But he also believes that research must underpin all eradication efforts, and that findings from the recent publication of the malaria parasite's genome and that of its vector, the mosquito, must quickly translate into new tools (see *Nature* **419**, 426–428; 2002).

The MIM aims to help build a vibrant African research community to scale up the battle against the disease. But an external evaluation of the initiative that was made public at the meeting says that there should be greater efforts to train African scientists and to improve links between researchers there and their colleagues in rich countries.

The MIM's funding of 36 new multicentre projects in Africa has empowered African labs, says Kojo Koram of the Noguchi Memorial Institute for Medical Research in Accra, Ghana, one of the continent's leading malaria centres.

Grants from MIM have also allowed Hassan Mshinda, head of the Ifakara Health Research and Development Centre in Tanzania, to set up labs for studying molecu-

lar markers of drug resistance, allowing country-wide 'early warning' maps of its occurrence to be drawn up.

The MIM evaluation suggests that its secretariat, which is now in Sweden, should eventually rotate among African countries, and be supervised by an advisory board with strong African participation. African involvement is fundamental, speakers at the meeting said, because even if money is forthcoming, it will achieve little unless the African countries make malaria a higher priority.

Many speakers also demanded that African states be held to account on the promises they made at Abuja. Only a handful have increased funding and staff to anywhere near the required levels. And only 17 African countries have reduced or lifted taxation on essential items for malaria control, while 26 are still taxing bednets. "No one is monitoring the Abuja claims," says Mshinda.

Many African researchers nonetheless remain optimistic. "Things are falling into place," said one of the scientists at the meeting. "We haven't seen this level of activity for decades." One area where substantial progress is being made is in the emergence of regional scientific networks, which are now generating continent-wide maps of baseline data on factors such as morbidity and mortality. "Success is inevitable," says Samba. "We will beat malaria, just as we beat river blindness." ■

C. DUFKA/REUTERS

US goes underground in scheme to bury greenhouse gas

Washington A test well some 3 kilometres deep is to be drilled on the border of Ohio and West Virginia as part of the US government's research into the possibility of storing carbon dioxide underground. In the new \$4-million project, researchers will investigate whether the gas could be pumped down into brine-filled formations. The well will be drilled at the site of a power plant in New Haven, West Virginia, owned by American Electric Power.

Researchers will determine whether the site is suitable for CO₂ storage before any decision is made about pumping gas into the formations. Several sites around the world are already being used to store carbon dioxide in this way to reduce greenhouse-gas emissions, most notably in the Sleipner gas field under the North Sea (see *Nature* **411**, 518; 2001).

The Ohio–West Virginia scheme was announced on 21 November, as the Department of Energy revealed plans for a national network of regional CO₂-sequestration partnerships between government, industry and universities. Environmental groups have criticized such

schemes, saying that they divert attention away from the need to cut greenhouse-gas emissions.

Drive to eradicate polio targets Africa and India

London Wild polio virus circulating in India, Nigeria and Egypt represents the biggest challenge to the World Health Organization's efforts to eradicate the disease worldwide by its 2003 deadline, an expert panel announced last week.

The Global Technical Consultative Group for Poliomyelitis Eradication, the campaign's independent auditor, said that at least six rounds of "high-quality" immunization are needed in the first half of 2003 in India,



Intensive care: a concerted immunization drive is needed to eradicate polio by the end of 2003.

Nigeria and Egypt if polio is to be eradicated by the end of the year.

But the group stressed that solid progress was being made, and that the wild virus is now present in limited areas of just seven countries worldwide, down from 10 at the beginning of 2002.

Vaccination programmes in Bangladesh, the Democratic Republic of the Congo, Ethiopia and Sudan should stamp out the virus there by the end of next year. A rapid injection of cash is also required to fill the programme's US\$275 million funding gap, \$50 million of which is needed for the first six months of 2003.

Stanford gets the green from energy firms

San Francisco Stanford University is to receive \$225 million over ten years from industry for studies aimed at cutting greenhouse-gas emissions. The donation equals the total amount given by corporations for research at Stanford over the past ten years.

The energy companies ExxonMobil and General Electric, together with the technology-services company Schlumberger, will invest the money in the project over the next decade. Stanford's Global Climate and Energy Project will

include studies of renewable energy sources, such as biomass and wind power, as well as new nuclear power technology. Low-emission transportation systems will also be investigated.

Franklin Orr will step down as dean of Stanford's School of Earth Sciences to run the project. He predicted that about half of the research will be at Stanford, with the rest taking place at other institutions under the guidance of Stanford scientists.

♦ <http://gcep.stanford.edu>

Britain cowed as scrapie-resistant sheep get BSE

London Sheep that are resistant to scrapie can, in extreme circumstances, become infected with bovine spongiform encephalopathy (BSE), according to research that is being carried out by Britain's Department for Environment, Food and Rural Affairs (DEFRA).

The study at the Institute for Animal Health in Compton, Berkshire, was part of a plan to replace British sheep with varieties resistant to scrapie, a prion disease that affects much of the country's stock. It was hoped that the scrapie-resistant sheep would be protected against BSE, which is also a prion disease. But one of the 19 animals in the study developed BSE-like symptoms 33

months after infected material was injected into its brain. Those sheep that received the material orally did not succumb to the disease.

Sheep with low scrapie resistance developed symptoms of BSE more rapidly after injection, and also developed the disease after receiving material orally. A DEFRA spokesman says that the department will continue with its restocking plans, although the animals are not totally resistant to BSE.

IBM gets up to speed on supercomputer plan

Washington The US Department of Energy (DOE) has laid plans for what it says will be the two fastest supercomputers in the world. IBM will receive \$290 million to build 'ASCI Purple' and 'Blue Gene/L', which will be housed at the Lawrence Livermore National Laboratory in California.

ASCI Purple, which will have a peak performance of 100 trillion calculations per second, will be used to run simulations of nuclear tests for the DOE's Stockpile Stewardship Program to assess the reliability of US nuclear weapons without conducting real tests. Blue Gene/L, which will handle up to 360 trillion calculations per second, is designed to evaluate the use of new computer architectures in stewardship computing.

Public inquiry considers Cambridge primate centre

London The latest round in a bitter dispute over plans to build a neuroscience research facility at the University of Cambridge, UK, kicks off this week, with the opening on 26 November of a public inquiry.

Permission for the centre has been rejected twice by the local council, most recently when police warned that it could attract violent protests (see *Nature* 415, 725; 2002).

The centre would use primates in research into conditions including Alzheimer's and Parkinson's disease. The British government has publicly backed the centre, but animal-rights groups are opposed to it.



Protest note: police warned that the primate lab could be targeted by animal-rights protesters.

M. FEARNS/PA



Wanna bet?

Scientific wagers have a long and colourful history. Are they just harmless fun, or can they help to frame and clarify important issues? Jim Giles surveys the odds.

Stephen was 33 when he made his first bet — an innocent wager with a colleague, just a token prize and professional pride at stake. It was not to be his last. Stephen's career went from strength to strength, but he continued to place wagers. When it comes to making a point about science, Stephen Hawking is a compulsive gambler.

The world's best-known cosmologist is not alone. The history of science is littered with bets, from ill-fated attempts to prove that the world is flat, to numerous wagers over whether various sub-atomic particles exist. A book at the Stanford Linear Accelerator Center in California, for example, records about 35 bets in high-energy physics dating back to the 1980s, many still unresolved. And Cold Spring Harbor Laboratory in New York is running Genesweep, a sweepstake on the number of genes in the human genome.

Genesweep's winner will pocket at least \$750, and gain the satisfaction of having out-guessed a star-studded cast of biologists. But making bets on science has a serious side. By putting their hard-earned cash on the line, wagers encourage scientists to think hard about their arguments and can also attract media attention to otherwise arcane topics. Over the past few years, websites have sprung up to harness these benefits. "A well-conceived bet can frame an issue," claims Kevin Kelly, editor-at-large of *Wired* magazine, and co-founder of the Long Bets Foundation, which runs one such website. "If it has

enough clarity, it can move the subject on."

But scientific wagers have a history of stirring controversy. Take the 1980 challenge laid down by the late economist Julian Simon of the University of Maryland, College Park. Annoyed by claims from environmentalists about the scarcity of natural resources, Simon asserted that the price of five metals would fall by 1990, and challenged dissenters to a wager. Winning the bet, Simon reasoned, would show that these resources are becoming more plentiful, not less so.

Paul Ehrlich, a population biologist at

Stanford University in California, took Simon on. Together with John Harte and John Holdren, physicists then both at the University of California, Berkeley, Ehrlich agreed to monitor the value of an imaginary portfolio for \$200 of each of the metals. If the portfolio's value dropped, Ehrlich, Harte and Holdren would pay Simon the difference; if it rose, he would pay them. The metals' value dropped by \$576 and Ehrlich and his colleagues duly paid up.

Although that bet was settled, the two sides disagree on the lessons to be learned. Some economists use the result to claim that environmental groups exaggerate the problems facing the planet. But Ehrlich maintains that it merely shows that the metals chosen were poor indicators of our exploitation of natural resources. "I regret entering the bet," he says.

Other scientific gamblers have emerged victorious and yet still suffered regrets. In 1870, the British naturalist Alfred Russel Wallace, co-originator of the theory of evolution by natural selection, took up a challenge to prove that the Earth is not flat. John Hampden, a dedicated flat-Earther, had staked £500 on the question — then a great deal of money. A test, involving a stretch of the Old Bedford Canal, north of London, was agreed on.

Wallace measured the canal's curvature using two markers, separated by about five kilometres and suspended at equal heights above the water's surface. Viewed through a telescope mounted at the same height some 10 km away from the farthest marker, the nearest



Brief wagers on time: Stephen Hawking is keen to bet on unanswered questions in physics.

one appeared to be the higher of the two. An independent referee agreed that this showed the Earth's surface to curve away from the telescope, and Wallace received his money.

But Hampden never accepted the result, and bombarded Wallace and his associates with abuse. In June 1871, he wrote to the naturalist's wife: "If your infernal thief of a husband is brought home some day ... with every bone in his head smashed to a pulp, you will know the reason." Wallace brought and won libel suits, but Hampden was declared bankrupt, leaving Wallace to pay costs. In the end, his winnings were wiped out — and the publicity attracted new members to the flat-Earth movement.

Other wagers have produced more fruitful debate. In 2000, Steven Austad of the University of Idaho in Moscow bet Jay Olshansky, a fellow researcher of ageing at the University of Chicago, that someone alive at that time would live to be 150, with their cognitive faculties intact. The two established a trust fund that they estimate will be worth \$500 million when the bet pays out to one or the other's heirs in 2150.

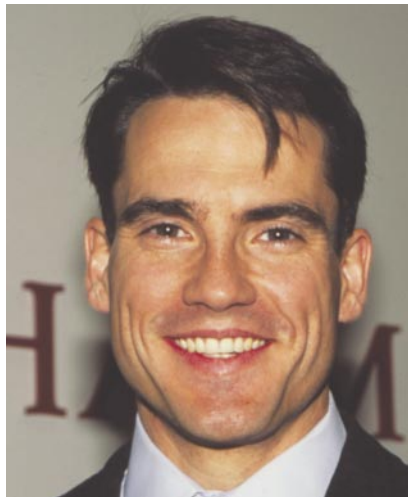
The wager has been widely discussed in the media. "Bets can be an interesting way to popularize science, as long as those involved make sure that they discuss science when reporters come calling," says Olshansky. He and Austad have been invited to speak in December before the President's Council on Bioethics in Washington. "If our wager contributed to the decision to have this discussion, then it was extraordinarily productive," Olshansky says.

Brought to book

Hawking's bets have similarly generated media interest — most recently, journalists have picked up on a \$100 bet made in December 2000 with theoretical physicist Gordon Kane of the University of Michigan in Ann Arbor. Kane asserts that the Higgs boson, the predicted particle thought to give other particles their mass, will be discovered at the Tevatron accelerator at Fermilab near Chicago; Hawking says it won't.

So could such publicity, and the way in which bets force both parties to hone their arguments, be harnessed in an organized way? The backers of the Long Bets Foundation think so. The foundation, established last year, aims to encourage people to propose long-term, well-defined bets and to generate discussion about the issues involved. Eleven wagers, worth a total of \$48,000, have so far been agreed on the foundation's website.

Other bets remain open. Bruce Damer, president of DigitalSpace, an Internet company in Santa Cruz, California, is looking for someone to challenge his assertion that "by 2024, 'artificial' life emerging somewhere out of the soup of human technology will be given a Latin taxonomic name by biologists ... and declared viable for study". Damer has staked \$1,000, the minimum



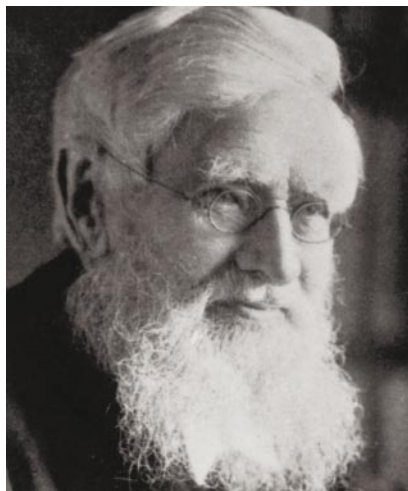
Futures trader: Tom Bell wants to assess scientific theories by having experts invest in 'idea stocks'.

allowed. This is kept in a trust fund and the money generated will be paid to the winner's chosen charity when the bet is resolved.

Kelly believes that the rules encourage responsible predictions. "We feel that there should be some pain in losing," he says. "Normally there is no penalty for being wrong, so people make predictions that are not responsible."

Other groups are trying to develop these ideas further. In the early 1990s, Robin Hanson, an economist now at George Mason University in Fairfax, Virginia, developed the concept of 'ideas futures' — markets that trade shares in ideas. Several versions of his concept, such as the web-based Foresight Exchange, which allows players to trade pretend-money shares in ideas, have been implemented by enthusiasts for market-based solutions.

You might, for example, want to assert that an adult human will have been cloned by 2005. You would start by 'buying' pairs of coupons in the idea from the market. The



Alfred Russel Wallace won his bet that the Earth is round, but was vilified for years by his opponent.

'YES' coupon pays \$1 if the claim comes true; 'NO' coupons pay the same if it is false. Confident that a clone will be created, you would retain your YES coupons. Sceptics would buy your NO coupons, believing that they would generate a pay-off in 2005.

A problem shared

Importantly, the price at which other players are willing to buy coupons indicates how much faith the market has in the idea. The human-cloning claim already exists on the Foresight Exchange. YES coupons were trading at 36 cents as *Nature* went to press, indicating that the market believes there is a 36% chance of the claim being correct.

This evaluation may be of limited use, however, as anyone can play the Foresight Exchange and traders are not using real money. But what would happen if an ideas market were to be played by scientific experts using their own cash? According to Tom Bell, a lawyer at Chapman University in Orange, California, such a market would be extremely useful, as the price of shares in a particular idea would provide a snapshot of how the scientific community felt about that issue.

Consider a claim about climate change, such as the size of the expected rise in mean global temperature by 2100. As long as enough scientists with relevant knowledge played the market, the price should reflect the latest developments in climate research. Policy-makers could use the market as a way of assessing current thinking, free from the bias of industry and activist groups — both of which tend to quote temperature changes at the extreme ends of the spectrum.

Bell is now looking for an institution to host the project, which he has named the Simon Market in honour of Simon's work. But could scientists be persuaded to sink their money into such a market? "If you present it as a scientific experiment then scientists would be eager to try it," suggests Bell. "Scientists also have egos and like money as much as anyone else." His proposal is currently being considered by the Mercatus Center, a public-policy, law and economics offshoot of George Mason University.

Whether the ideas of Bell and the Long Bets Foundation will sway public debate or merely provide entertainment remains unclear. But while you're totting up the odds, there's still time for visitors to Cold Spring Harbor to enter Genesweep, before the final figure is announced next year. Our tip: 31,789 genes.

Jim Giles is *Nature's* associate News and Features editor.
Genesweep

♦ www.ensembl.org/Genesweep

Long Bets Foundation

♦ www.longbets.org

Foresight Exchange

♦ www.ideosphere.com

The Simon Market

♦ www.simonmarket.org

Open the floodgates!

Hydrologists are gearing up for a second attempt to restore the altered environment of the Grand Canyon, by letting the Colorado River run free. Kendall Powell discovers the lessons learned from the first, fruitless flood.

On 27 March 1996, David Topping woke up at his campsite in the upper Grand Canyon to see a rushing Colorado River that ran three metres higher than the previous day. The flood continued for seven days, as 1,290 cubic metres of water per second flowed through the bypass tubes of the Glen Canyon Dam, more than 160 kilometres upstream.

As the torrent churned up sediment, Topping recalls, the crystal-clear river turned to “the colour of chocolate milk”. This was just what Topping, a hydrologist with the US Geological Survey (USGS) Grand Canyon Monitoring and Research Center in Flagstaff, Arizona, wanted to see: the goal was to rebuild sandbars and beaches that had been eroded since the dam was completed in 1963. The flood, Topping and his colleagues reasoned, should liberate years of accumulated sand, silt and clay from the river channel and deposit it onto the eroded beaches.

But David Rubin, a USGS sedimentologist at the Pacific Science Center in Santa Cruz, California, soon began to suspect that something was amiss. From his vantage point 56 kilometres downstream from Topping, the chocolate-milk colour soon faded, although the flood continued to rage. “Every day the water got a little clearer while the flow stayed the same,” he says.

Today, Topping, Rubin and their colleagues know that a chronic lack of sediment in the river downstream of the dam doomed the flood to fail in its goal of restoring the Grand Canyon’s beaches to a more natural state. Armed with this knowledge, they are now planning to flood the canyon once more, timing the deluge to coincide with an abundance of fresh sediment.

This ambitious hydrological experiment



is part of a wide-ranging plan to preserve the canyon’s unique geomorphology and ecology — including a rescue plan for the endangered humpback chub (*Gila cypha*). For the scientists involved, it is also an experiment in stakeholder relations: power companies, tourist operators and Native American tribes all have vested interests in the outcome, and their views must be taken into account.

Grains of hope

With so many people to please, project scientists desperately wanted the 1996 flood to be successful. And at first, optimism ran high. Aerial photos showed wide stretches of sandy beach at more than 50 points down the Grand Canyon. Officials with the federal Bureau of Reclamation, which manages water resources in the western United States, were quick to declare the experiment a success, arguing that the redistribution of sediment would restore fish habitats and make rafting and camping within the canyon more enjoyable.

The day after the floodwaters receded, however, Rubin began analyses that would give the lie to this rosy picture. The first

trench he dug through one of the renewed sandbars revealed an odd pattern. As a deposit is built during a flood and its subsequent recession, heavier sand particles would be expected to settle out first, followed by fine silt particles that would settle as the water slowed again. But instead, Rubin found coarse, large-grained sand at the top of the trench, with silt at the bottom.

Two months later, Rubin returned to the canyon with Jack Schmidt, a geomorphologist at Utah State University in Logan, and found the same pattern at several dozen points along a 220-kilometre stretch of the river¹. In the months that followed, the restored beaches quickly eroded away. By the time of the flood’s first anniversary, there was little to show for the exercise.

We now know why. Before the flood, project scientists had asked whether there was enough sediment in the river for the torrent to cause a net benefit. “Our best estimate at the time was that we had enough,” says Schmidt. But the experts were sadly mistaken.

What Rubin had spied during the flood was confirmed by analyses of sediment in floodwater samples². These revealed that the sediment supply had diminished rapidly with each passing day, while the average size of the particles increased. This explained the large particle sizes at the surface of Rubin’s trenches.

Topographical studies showed that the flood had actually reduced the volume of sandbars in the upper canyon, providing a clue as to where this coarse sand had come



Grand plan: Jack Schmidt and his colleagues believe they now have the knowledge to ensure that a second flood project is successful.

J. SCHMIDT

M. COLLIER

from^{3,4}. Floodwater samples also revealed that most of the suspended sand matched the type found in these sandbars. The restored beaches, it transpired, were built not from sediment scoured from the river channel and deposited as the water receded, but from sand scooped up from the sandbars that were the beaches' own foundations — hence the ease with which the recreated beaches were subsequently eroded.

Further analyses have shown that the flood formed the majority of beaches within its first two days⁵. Indeed, as the flood reached its end, it was eroding the beaches rather than building them up. Subsequent studies have revealed why the supply of sediment was so limited: most of the sand pumped in from the river's tributaries after seasonal rains is transported downstream in a matter of weeks or months, rather than being retained for several years, as researchers had thought^{6,7}.

Armed with this knowledge, Rubin, Topping, Schmidt and their colleagues have devised a new approach for future flooding experiments⁸. To rebuild beaches properly, the scientists believe, floods should ideally be timed to coincide with fresh sediment input from the Colorado's tributaries. Seasonal rains in late summer and autumn transport roughly three million tonnes of sediment each year down the Paria River, the main tributary that enters the Colorado upstream of the Grand Canyon (see map, opposite). If a flood were to be let loose from the Glen Canyon Dam after this input of sediment, just 60 hours of flooding, rather than the seven days tried previously, should be sufficient to build beaches that will be much more resilient.

Using these recommendations, the Bureau of Reclamation, the USGS and the National Park Service have crafted a new set of experiments for the Grand Canyon under the watchful eye of the Glen Canyon Dam Adaptive Management Workgroup — a diverse group of stakeholders that represents the companies that supply power generated by the dam, the governments of the states along the Colorado River, park managers, Native American tribes, recreational interests and environmental groups.

Getting all these factions to back the project is no easy task, particularly given that the



What a blast: the Glen Canyon Dam during the 1996 effort to restore the Colorado River's beaches.

1996 flood cost the hydroelectric companies that use the Glen Canyon Dam's power some \$2.5 million in lost generation. This, together with the experiment's disappointing results, prompted complaints from some workgroup members. But the scientists involved argue that the first flood yielded valuable lessons. "As an experiment, it was a tremendous success," says Schmidt. "We learned important things we could not have known had the experiment not been run — that much less sediment than we thought can be stored on the river bed, and that the sediment budget used prior to 1996 had been calculated in error."

Don't rock the boat

The Glen Canyon Adaptive Management Workgroup has to balance a diverse range of interests. Rafting guides and National Park rangers want the best experience for boaters and campers without having fluctuations in flow disrupt their activities. The Western Area Power Administration, which sells the dam's hydroelectric power to companies in 15 states, must meet the power demands of its customers. Archaeological sites, plants of tribal cultural importance, and threatened species must also all be protected.

These goals often conflict with the aim of restoring the canyon to its natural state, and any management plan will contain inherent uncertainties. But through regular meetings

and reviews of the scientific evidence, the stakeholders represented in the programme have recommended to the Secretary of the Interior, Gale Norton, that the next round of experimental flooding should go ahead. Norton has the final say for the dam's operations and the management of the Grand Canyon National Park, and her decision on how to proceed is expected any day now.

Although the various stakeholders have diverging priorities, everyone agrees that, since the Glen Canyon Dam became the Grand Canyon's gatekeeper almost four decades ago, the canyon and its river have undergone a detrimental transformation.

Before 1963, the Colorado's flows varied seasonally and carried an average of 57 million tonnes of sediment each year, giving its waters a muddy reddish colour. In the spring, snowmelt from the Rocky Mountains would bring a rush of near-freezing water. But as summer wore on, the river would slow to a lazy creek and warm to as much as 32 °C. The spring snowmelt floods, which could reach heights more than twice that of the 1996 experimental torrent, scoured sediment from the channel and renewed the Grand Canyon's beaches.

Today, flows remain relatively low all year round. With more than 90% of the river's incoming sediment trapped behind the dam in Lake Powell, the water flowing through the

T. SMART/DESERET NEWS



news feature



Coming out in the wash: the 1996 flood restored the Colorado River's eroded beaches (left, before; right, after), but the benefits were sadly short-lived.

Grand Canyon is unnaturally clear. Deprived of its regular floods, the river has lost the ability to regenerate its shores. And because water flows from the dam through valves near its base, insulated from the Sun's rays, the river's temperature remains a chilly 9 °C throughout the year.

These changes have been bad news for the Grand Canyon's native fish populations. Three species have already disappeared from its waters. The humpback chub, which evolved in the Colorado River basin in the past two million years, and is found nowhere else, has declined from an estimated population of 8,000 to around 2,000 over the past 20 years.

The chub is a strange-looking fish. Before the arrival of the dam, its prominent hump probably stabilized its swimming through the river's turbulent, fluctuating flows. Eyes were fairly useless in the canyon's muddy waters, so they shrank to two small dots. But this evolutionary oddball has done poorly in today's cool, clear waters. "The chub has done such an amazing job of surviving in the naturally harsh desert river system," says Nikolai Ramsey, a programme manager at the Grand Canyon Trust, a conservation group in Flagstaff. "The irony is that it can't survive in these really stable conditions."

In part, the chub's decline is due to the loss of beaches and sandbars, which provided sheltered backwaters for its fry. At the same time, populations of brown trout (*Salmo trutta*) and rainbow trout (*Oncorhynchus mykiss*) in the canyon have exploded to 250,000 or more, constituting a predatory threat. In a 1997 study, biologists retrieved partially digested chub from the stomachs of trout and channel catfish (*Ictalurus punctatus*), another non-native species⁹. Some researchers suspect that the trout are also feeding on chub eggs and fry. "Even if there were suitable backwater habitats for chub, they wouldn't stand a chance," says Steven Gloss, a biologist at the USGS Grand Canyon Monitoring and Research Center in Flagstaff. "We've got to get the numbers of trout down."

To do so, the USGS will take a two-pronged approach over the next two years. First, it will remove non-native fish from a 16-kilometre stretch of the Colorado River in the canyon, near chub spawning grounds, by electrofishing. Wildlife managers will

collect the stunned fish and kill the non-natives with an overdose of anaesthetic. The native fish will be allowed to recover in river water before being returned.

The second part of the plan is to disrupt trout breeding by varying the flow from the Glen Canyon Dam during the spawning season, from January to March. Because mating occurs in the shallows of the river, alternating between high and low flows should leave trout eggs and fry stranded.

Local fishing guides say that it is inevitable that some adult trout will also be left high and dry. The government acknowledges that this is a risk, but argues that the plan will nevertheless benefit anglers. Today, the fishery is clogged with many small trout, which compete with each other. Thinning the population will result in larger trophy fish, argues Gloss.

Fishing for compromise

The local Hopi and Hualapai tribes are also anxious about the trout-control plan. "We have an ancient belief that the fish are our ancestors and they are very important, regardless of whether they are an exotic species," says Loretta Jackson, cultural resource manager for the Hualapai. But the tribes are willing to consider a compromise plan in which the killed fish won't simply be discarded, but will be put to beneficial use, such as for pet food or fertilizer.

So far, the Glen Canyon Adaptive Management Workgroup has avoided show-stopping controversies. In general, most participants agree that the compromises reached, although laborious, are a better alternative than a courtroom standoff. But litigation remains a possibility, particularly from environmental groups, which remain frustrated by the slow progress in protecting the humpback chub. "We'll stay in the group



Getting the hump: the humpback chub is under threat from habitat change and predatory fish.

and work sincerely within it — as long as it's our best option," says Ramsey.

In this strained atmosphere, scientists are awaiting the green light for the next set of experiments. The revised plan calls for a flood of similar magnitude to 1996, but only after a significant influx of sediment from the Paria River and only for two and a half days. Even if Norton were to give the go-ahead tomorrow, however, another necessary compromise means that the next flood can't take place until after 1 January, because of laws that govern the dam's operations. To compensate for this, the plan calls for low flows from the Glen Canyon Dam following a large sediment input, to try to retain as much of it as possible in the river channel. Scientifically, it's a tenuous position; realistically, it's the only current option.

Mother Nature has not played her part this year — not enough sediment has entered the river to put the flood plans into motion even if they are approved, because 2002 has been one of the driest years on record. But another bad drought year is historically unlikely, and the plans will carry over to next year. The removal of non-native fish will proceed regardless.

In the meantime, the scientists advising the adaptive management process face a heavy responsibility. "We have river guides, power customers and environmentalists looking over our shoulders. It's a pain in the neck, time-consuming; collecting data is difficult and expensive," says Schmidt. "But it's one of the most beautiful places on Earth. I would hope people would care about it." ■

Kendall Powell is an intern with *Nature* in Washington.

- Rubin, D. M., Nelson, J. M. & Topping, D. J. *Geology* **26**, 99–102 (1998).
- Topping, D. J. et al. in *The Controlled Flood in Grand Canyon Geophysical Monograph* 110 (eds Webb, R. H. et al.) 71–98 (Am. Geophys. Union, Washington DC, 1999).
- Hazel, J. E. Jr et al. in *The Controlled Flood in Grand Canyon Geophysical Monograph* 110 (eds Webb, R. H. et al.) 161–184 (Am. Geophys. Union, Washington DC, 1999).
- Andrews, E. D. et al. in *The Controlled Flood in Grand Canyon Geophysical Monograph* 110 (eds Webb, R. H. et al.) 117–130 (Am. Geophys. Union, Washington DC, 1999).
- Schmidt, J. C. in *The Controlled Flood in Grand Canyon Geophysical Monograph* 110 (eds Webb, R. H. et al.) 329–342 (Am. Geophys. Union, Washington DC, 1999).
- Topping, D. J., Rubin, D. M. & Vierra, L. E. Jr *Wat. Resour. Res.* **36**, 515–542 (2000).
- Topping, D. J. et al. *Wat. Resour. Res.* **36**, 543–570 (2000).
- Rubin, D. M. et al. *Eos* **83**, 273, 277–278 (2002).
- Marsh, P. C. & Douglas, M. E. *Trans. Am. Fish. Soc.* **126**, 343–346 (1997).

Could a college merger close the transatlantic gap?

Why UCL, home of Britain's most-cited scientists, should merge with third-place Imperial.

Sir — In your Opinion article (*Nature* **419**, 763; 2002) about the benefits of a merger between Imperial College and University College London (UCL), you voice concern that these two London colleges will dominate the UK research scene, and you comment that this situation does not happen in the United States. Similar concerns led a group of UK academics and students to vociferously oppose the move, resulting in newspaper reports that the merger has been postponed. I contend that this largely London-based community will come, in time, to be viewed as dangerously parochial. It is precisely the international context, which your Opinion article alluded to, that renders the UCL–Imperial merger so imperative right now, if Europe is to remain academically competitive with the United States (see, for example, Wilhelm Krull's Commentary, *Nature* **419**, 249–250; 2002).

According to the ISI (www.isihighlycited.com) list of the world's 1,200 most cited

scientists, the gap between the United States and the rest of the world is widening at an alarming rate. Only about 80 of those scientists are in the United Kingdom. In microbiology, for example, eight of the world's leading scientists are at Harvard Medical School, 71 elsewhere in the United States, and only six in the United Kingdom. Nevertheless, some UK universities do attract clusters of the world's top scientists. Top of the list is UCL with 12, Cambridge is second with 10 and Imperial is third with six. Oxford has only three.

A merger between Imperial and UCL would allow the creation of an institution that for the first time could have the clout to shake up the UK scene. It might allow universities to break free from the constraints of uniform pay, and ensure that academics here are at last rewarded in ways that will stop them bailing out to the United States and elsewhere.

Over and over again, reports about higher education in the UK national press

imply that there are only two institutions that can nurture talent: Oxford and Cambridge. This myopia is very damaging, particularly as it limits the interest of overseas students and academics in UK-based research.

The ISI list indicates that there is a lag in public understanding of who really is at the very top of UK higher education, which Imperial and UCL's proposed merger tackled head-on. The publicity surrounding their proposal has already done much to restore the reputations of both institutions to their rightful place.

The way to bring this issue to public and international attention with the urgency it deserves is to merge universities, producing institutions tremendous enough to demand an entry into the academic 'World Series', and to show that the tournament doesn't just belong in the United States.

Raj Persaud

Alumni Society, University College London,
Gower Street, London WC1E 6BT, UK

Unforeseen growth of academic astrology

Sir — I was appalled when I read a report in *Nature* last year about proposals by the Indian government to encourage universities to teach astrology (*Nature* **411**, 227; 2001), but never thought that anything similar might occur in the West. It now seems that I was mistaken.

A British organization called the Sophia Project declares on its website (www.sophia-project.org.uk) that its goal is "to advance the scholarly study of astrology and cultural astronomy in British institutions of higher education". The Sophia Project currently funds four principal initiatives at British universities (the Warburg Institute at London; Southampton; Kent at Canterbury; and Bath Spa University College). The initiative at the University of Southampton, which is called the Research Group for the Critical Study of Astrology, states on its website (www.rgcsa.org) that its disciplinary scope includes "research in the field of psychology (ably begun by the late Hans Eysenck) on the links, if any, between personality and more complex indicators of supposed planetary influence".

One research project being funded in this way is a study of the possible impact of the planets on fertility and childbirth; another is the possible correlation of birth date and prostitution; and another

is the effect of Jupiter on alcoholics.

The Sophia Project, which funds the Southampton group, is funded by a private benefactor, but the participation of any institution of higher education surely entails the use of at least some public money in support of bogus research.

Dylan Evans

Department of Mechanical Engineering,
University of Bath, Bath BA2 7AY, UK

DNA database could end problem of identity fraud

Sir — Although no system is perfect, your correspondents (*Nature* **419**, 247–248; 2002) seem to have missed the point about DNA testing, made by R. Williamson and R. Duncan in their Commentary "DNA testing for all" (*Nature* **418**, 585–586; 2002). First, it is the most reliable method of identification. Second, DNA evidence has brought about more acquittals on appeal after false convictions than convictions of the innocent. Third, identity theft and identity fraud are bigger problems for most citizens than a theoretical risk of false conviction or state coercion (for which a DNA database is hardly a necessity), not to mention the issues of preventing or dealing with mass fatalities like those of 11 September 2001.

Forensic databases are arguably discriminatory in that only suspects and

those previously convicted are in danger of falling victim to a false match. A universal database would be equally fair or unfair to everyone.

The UK government introduced a requirement to register births, deaths and marriages in 1837. To implement a twenty-first century register backed by DNA is simply modernization.

Martin Evison

Forensic Pathology, Sheffield University Medico-legal Centre, Sheffield S3 7ES, UK

Caption confusion

Sir — The photograph illustrating your News story "Suspensions intensify over elusive European Academy of Sciences" (*Nature* **419**, 855; 2002) is not of Brussels but of another Belgian city: Bruges. This city, called "the Venice of the North" because of its many canals, is well known to tourists. I know nothing about the organization discussed in your article, but what does exist in Bruges is the European College (www.coleurop.be), a post-graduate institute of European studies.

Jean-Paul Schiepers

J.P.S@skynet.be, Belgium

The caption supplied by the photo library to accompany this picture identified the city as Brussels. *Nature* regrets this error, which was pointed out by several readers — Editor, Correspondence

No more moa

Human activities depleted much of New Zealand's fauna. Let it be a warning.

The Lost World of the Moa:
Prehistoric Life of New Zealand
 by Trevor H. Worthy &
 Richard N. Holdaway
Indiana University Press: 2002. 718 pp.
\$89.95, £57
Stuart Pimm

Ninety years ago, Urupeni Puhara, a nonagenarian Maori chief, said that 'te kura' ('red bird'), not 'moa', was his ancestors' name for the country's great bird. 'Moa' means chicken, and its application, the chief thought, was a mistake by the Victorians who first found the giant bones. Neither he nor his grandfather had seen the bird, its eggs or its footprints. He heard stories that the bird had lived all over the North Island, but that it disappeared "after the coming of Tamatea", who set fire to the land.

This book elaborates on these stories, documenting in extraordinary detail what we now know about the fate of New Zealand's fauna and landscapes before and during the first human settlement some 700 years ago. Although there were regional changes in the distribution of New Zealand species at the end of the Pleistocene, there were no extinctions until after first contact. Then, armed with only limited Stone Age technologies (no bows or specialized spears, but certainly fire), Tamatea and his band of settlers extensively burned the forests and eliminated over 60 species of bird found nowhere else on Earth. We do not even know these birds' names.

Moa bones informed nineteenth-century debates, as they do modern ones. A missionary sent bones to England, where Richard Owen wrote about them in 1840. Extinction was then a relatively new idea, a denial of Scripture that required either a multiplicity of floods and creations or the hope of finding the species alive somewhere. Moa were so large that they would be hard to miss, and their demise was the subject of much debate. Some blamed climate change, and others felt that moa were in a kind of inexplicable but terminal decline before humans arrived.

This book's best chapter is its last, relatively short one. It summarizes the merits of these alternatives with 500 pages of detail in hand. The 'terminal decline' hypothesis falls to the mounting evidence that moa persisted until first contact, but it is also generally unscientific. The hypothesis states that the species that were lost were somehow flawed, but neither explains those flaws nor predicts which species will be next.

The 'climate change' hypothesis becomes increasingly implausible as more accurate

methods date the extinctions. In the past 700 years, New Zealand's climate has changed little. Earlier changes were substantial, but caused no extinctions. The evidence that birds were slaughtered for food is so abundant that bones from some archaeological sites were hauled off by the wagon-load for use in fertilizers. The most telling evidence is that only early settlements contain moa bones. Moa bones were "rare to nonexistent" in the diet of former moa hunters within the century after first settlement.

Hunting was not the only mechanism involved. Introduced rats probably eliminated many ground-nesting birds, including seabirds. Many seabird species became extinct after first contact, putting the onus on the climate-change hypothesis to explain marine as well as terrestrial changes. As Maori legends recount, and as radiocarbon dates confirm, fires changed the landscape rapidly and extensively. They probably destroyed half of the natural vegetation of the South Island. But Tamatea's band was only the first wave of human settlers. Following James Cook's mapping of the islands, contact with Europeans further changed the islands' habitats and diminished their diversity.

The consequences of both contacts were the loss of 41% of the islands' endemic bird species. Some 49 (51%) of the North Island's bird species have gone, along with 53 (47%) of the South Island's, and another 30 species are endangered, many of which survive only with active and expensive management. New Zealand's fauna only survived as well as it did because small offshore islands provided the last refuge for many species.

Hypotheses and statistics do not properly capture why this book is so compelling. It is not just about moa. Its core is an extensive description of kiwis and waterbirds, a huge eagle (one of the heaviest flying birds), the islands' many species of rail, its shorebirds, parrots and penguins, along with frogs, bats and reptiles. One of the bats belongs to an endemic family, and one of the reptiles, the tuatara, has distinct ancestors back in the Jurassic period. New Zealand is no upstart volcanic island colonized from afar. It broke off Gondwanaland 80 million years ago, carrying with it a unique and ancient flora

and fauna. There was a unique coevolution too: the flora has an unusual number of divaricating plants, in which the branching shoots and twigs interlace. Life is different where the browsers are moa, not mammals.

Trevor Worthy and Richard Holdaway describe an isolated world that is an unfortunate model for present-day effects of human activities, where forest destruction is driving so many (overwhelmingly unnamed) species to extinction. Chief Puhara's oral tradition of fires, forest clearing and lost species contains deep and accurate insights into present-day human impacts. This book should prompt those who believe that 'things are getting better' to explain just how future massive global changes driven by sophisticated technologies and massive energy consumption will harm so few species, given that a few people with primitive technologies eliminated so many in New Zealand.

Stuart Pimm is at the Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708, USA.



The 'moa', or 'te kura', was once widespread on New Zealand's North Island. But with the arrival of the first human settlers, the bird went into rapid, terminal decline.

THE NATURAL HISTORY MUSEUM, LONDON

Weird notions that drive science

Strange Matters: Undiscovered Ideas at the Frontiers of Space and Time

by Tom Siegfried

Joseph Henry Press: 2002. 224 pp. \$24.95

Marc Kamionkowski

For almost 30 years, physicists have been facing a long haul across the almost interminable desert to the grand unified theory (GUT) of strong, weak and electromagnetic forces. This theory postulates an almost unimaginably huge gap between the energies of current particle accelerators and those at which the manifestations of unification become most apparent. If that's not sufficiently daunting, string theorists have speculated further, suggesting that gravity is tied to the other three interactions at an even higher energy scale.

However, in the past few years some theorists have begun to question whether this desert is real. String theory has long postulated the existence of additional spatial dimensions, beyond the three that we know, but these have always been thought to be microscopic. It has recently been suggested that some of these dimensions may be as large as a millimetre. To see how a dimension can have size, imagine tightly rolling a sheet of paper; the resulting 'curled' dimension will be much smaller than the original dimension. If these dimensions are a millimetre in size, the extraordinary and mysterious weakness of gravity relative to the other three forces could be due to its leakage into other dimensions. If that's right, unification could be just around the corner. Instead of trekking across a vast featureless energy expanse, we may well see radical departures from physics as we know it in laboratory experiments any day now. These will tell us whether or not the large extra dimensions are just a theoretical mirage.

Both the GUT desert and these large extra dimensions are motivated, at least in part, by elegant mathematics. This 'prediscovery' of physical phenomena — their discovery in mathematical equations that anticipates their discovery in the laboratory — is the unifying thread of Tom Siegfried's enjoyable new book, *Strange Matters*. Although the eponymous matter here is most literally strange-quark matter, which is discussed in the opening chapter, other ideas, such as mirror matter, large extra dimensions, supersymmetry, dark matter, unification, string theory and even extra time dimensions are, by any reasonable standard, also pretty strange.

These subjects, as well as a variety of interspersed curious historical vignettes

(could Edgar Allan Poe really have inspired the physicist Alexander Friedmann to ponder an expanding Universe?), all serve to illustrate the intriguing power of mathematics in describing the physical Universe. The end result is not just a tour of current theoretical speculation on time, space and matter, but rather a cogent argument as to why we should take these exotic ideas seriously.

The consistent success of mathematical equations in describing physical phenomena has surprised and enticed scientists and philosophers for years. Siegfried's point is that mathematical physics is not just numerology or catalogues. There is something deeper, as can be seen in the numerous cases where scientists have been able to get more out of the equations than they put in. Paul Dirac's equations to describe the relativistic quantum behaviour of electrons required the existence of positrons — positively charged particles subsequently discovered by Carl Anderson. Friedmann's 'prediscovery' of the Universe's expansion presaged its observational discovery by Edwin Hubble. Wolfgang Pauli's neutrino, which was first required to satisfy energy conservation, was then found in particle beams.

Today an assortment of curious phenomena arise in the mathematics of theoretical physics. Which, if any, of these will be confirmed in the laboratory? Will it be

superstrings? Strange-quark matter? Large extra dimensions? Elementary-particle dark matter? Siegfried explains each of these questions, usually with clarity and integrity. Physicists will be gratified by his testimonial in support of the Big Bang, and impressed that he managed to work in an accurate description, accessible to the layman, of the branch of mathematics known as group theory.

The choice of topics is rather unusual, beginning with a search for strange-quark matter (something a bit off the beaten path of physics today) and then moving on to supersymmetry, string theory and large extra dimensions, which are central currents of modern particle theory. But this eclectic mix helps to set the book apart from other recent popular books on similar subjects. Although the prose is at times a bit dry, the pace is just right and the presentation engaging. Except for a few notable cases (such as a proposal that extrasolar planets are made of mirror matter), the discrepancy between the well established, plausible, unlikely and far-fetched is discerned clearly.

There is also some subtle humour. For example, at the end of the book Siegfried discusses duality — roughly speaking, this is the notion that several theories that seem

High noon at the meridian

Among the 5,000 scientific instruments at the National Maritime Museum at Greenwich in London, UK, is a remarkable collection of sundials and related instruments, ranging in origin from the Far East and Islamic countries to London itself.

It includes dipleidoscopes, like the one shown here, which were used to provide an accurate time check at noon, when the Sun passed over the meridian of the observer. The collection is catalogued in extraordinary detail in *Sundials at Greenwich: A Catalogue of the Sundials, Horary Quadrants and Nocturnals in the National Maritime Museum, Greenwich* (Oxford University Press, £99.50), edited by Hester Higton. The book also includes chapters that set the instruments in a cultural and historical context.



to describe different physical situations can actually be mathematical equivalents. This concept comes in handy when he tries to reconcile the apparent inconsistency in the thesis of this book — that great ideas drive progress in science — to that in his previous book, *The Bit and the Pendulum* (John Wiley, 2000), which postulates that technological advances drive scientific discovery. Although duality is not meant to imply that you can always have it both ways, in this case maybe he can. ■

Marc Kamionkowski is in the Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, California 91125, USA.

A social activist in genetics

**Making Genes, Making Waves:
A Social Activist in Science**

by Jon Beckwith

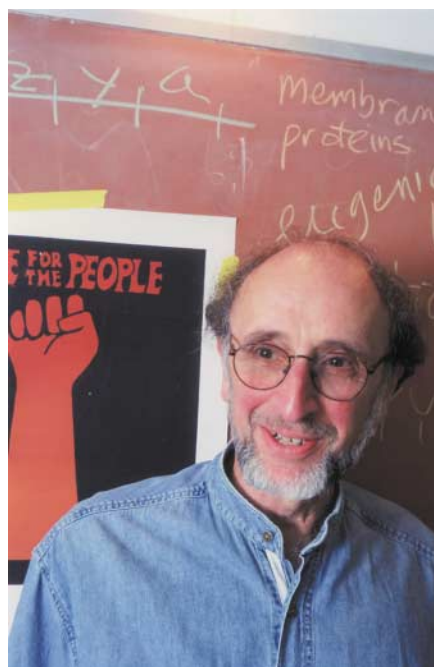
Harvard University Press: 2002. 256 pp.

\$27.95, £18.50, 27.95 euros

Ute Deichmann

“Science was puzzle-solving — figuring out mathematical proofs or devising pathways for the synthesis of complex organic compounds — it was fun.” The joys of doing science, first experienced in college, motivated Jon Beckwith to become a scientist — and to remain one. He is now professor of microbiology and molecular genetics at Harvard Medical School. But literature, philosophy and social concerns have remained important to him throughout his scientific life. In 1969, Beckwith was the first to isolate a single gene from a chromosome (of the bacterium *Escherichia coli*) and, two decades later, his experiments on protein secretion from cells opened up new lines of research on the process of protein folding.

In this beautifully written autobiography, Beckwith explains these and his other scientific successes, as well as his failures. He vividly describes aspects of the “cultural revolution in science that molecular biology brought with it”, epitomized by “iconoclastic and unchemist-like” Jim Watson, and major public controversies about genetics in the United States from the 1960s. Beckwith is acquainted with various laboratories in the United States and Europe, and characterizes the scientific styles of different individuals and groups. He was particularly fascinated with the French style of “daring leaps of logic”, simple experiments that seemed to yield profound insights, and papers with persuasive “elegant rhetorical strokes”, but later realized that it did not represent the process of how discoveries actually take place. He amply depicts the human elements, “the wrong turns, the surprises, the



Social conscience: Jon Beckwith was the first to isolate a gene, but warned of the risks of genetics.

flashes of intuition, even the passions that drive us in science”.

Beckwith's growing enchantment with science was mirrored by his growing concern for its consequences. His social activism in science grew out of a more general political radicalism in the 1950s and 1960s, stimulated especially by the civil-rights movement in the United States, the assassination of Martin Luther King, and the turmoil over the Vietnam War. As a member of the action group Science for the People, he was convinced that scientists have a special social responsibility, so he decided to help inform the public about the potential negative social consequences of genetic research. In 1969, in the same week that his famous paper about the first isolation of a gene appeared in *Nature* (224, 768–774), Beckwith called a press conference aimed at raising public awareness of the possible consequences of genetic manipulation.

This received huge international press coverage and contributed to rising fears, even among fellow scientists, about the possible dangers of molecular-biological research. But Beckwith fails to mention that most of the scientists who called for a moratorium on recombinant-DNA research in 1973, Watson and Paul Berg among them, later considered this a mistake and the fears unsubstantiated.

By contrast, Sydney Brenner, one of Beckwith's scientific heroes, never believed that scientists have a special social responsibility. In his autobiography, *A Life in Science* (BioMed Central, 2001), Brenner expressed the opinion that, in order to act responsibly, one should not prevent the generation of knowledge, but rather answer the following

question: “What are you doing with your knowledge once you get it?” This, of course, presupposes the neutrality of science, which Beckwith denies.

Beckwith's activism was also an outcome of his preoccupation with history. He realized that, in contrast to physicists, who had openly confronted their historical burden of the past, the atomic bomb, geneticists were ignorant of their own ‘atomic’ history — their role in the eugenics movement. Suspicious of genetic research that claimed to explain antisocial behaviour, he launched a campaign against a study at Harvard Medical School on the development of boys with an extra Y chromosome (XYY). Beckwith was concerned about the ethics of identifying and studying these children, because many people still believed previous, seriously flawed scientific claims that linked this chromosomal aberration with criminal behaviour. The Harvard researchers finally decided to stop the screening. However, this campaign, because of the distrust it caused between the activists and faculty members, affected Beckwith's life more than any other.

E. O. Wilson's book *Sociobiology: The New Synthesis* (Harvard University Press, 1975) presented a new theory about genetic programmes of behaviour in animals and humans. It received wide media coverage. In their public attack on the theory, in which they said it was biologically determinist, Beckwith and his colleagues went as far as to associate Wilson and his theory with Nazism. Whether the scientific evidence that Wilson presented was strong or weak, some of the attacks were blatantly unfair. Social activists in science have to show responsibility, too.

In 1989 Beckwith joined the programme to explore and anticipate the ethical, legal and social implications of the Human Genome Project. The resulting public discussion of this issue led to the passage of bills outlawing the practice in several US states. At the time, the antagonism between genome scientists and activists seemed to be unbridgeable, which Beckwith interprets as part of the long history of conflict between the world of science and the humanities, as described by C. P. Snow in his *Two Cultures* (Cambridge University Press, 1959). Today, bioethicists and genome scientists seem to have begun to close this gap.

Beckwith's account of social activism in genetics implies that it arose from more than an analysis of the possible dangers. It could also be predicted from scientists' political inclinations: the same group of leftist scientists in the United States was critical of such different issues as recombinant DNA and genetic theories of human behaviour. That molecular biology attracted more critical and socially active scientists than did older sciences such as chemistry does not mean that molecular biology is intrinsically more

dangerous than chemistry. Indeed, chemistry has its own 'atomic' history, which includes chemical warfare during the First World War, the use of Zyklon B in the Auschwitz concentration camp, and environmental disasters.

Beckwith has portrayed a fascinating period in the history of modern biology and

of the interaction of science and society in the Western world. Thanks to him and other activists, social injustices resulting from the application of genetics are now widely discussed and, in democracies, meet with legal measures and regulation. In this book Beckwith, a committed scientist — and here he

has many predecessors — calls for greater humility about what science can and cannot accomplish. This is a call that scientists would do well to take seriously. ■

Ute Deichmann is at the Institute of Genetics, Cologne University, Weyertal 121, 50931 Cologne, Germany.

Science in culture

The cut-and-paste carafe

Caravaggio's optical naturalism was a success in Rome in around 1600.

Martin Kemp

Caravaggio was a shocking painter. Or, rather, his compelling naturalism enthused a new breed of patrons and seduced young Roman artists, while scandalizing the artistic establishment. Previously, painters had spent long years mastering perspective and anatomy. Now a youthful provincial from Lombardy was achieving startling results without the necessary learning.

Caravaggio did not use drawings to map out perspectival forms within geometrically constructed spaces. Instead, he assembled vividly illuminated items in shallow spaces against dark backgrounds. His paintings appeared (like early photographs in the nineteenth century) to be acts of nature, rather than art.

A typical reaction was that of Giovanni Pietro Bellori, who singled out a painting of the carafe of flowers which had first brought Caravaggio to the attention of Cardinal Francesca Maria del Monte. The painter had miraculously captured "the transparencies of the water and... reflections of the window of a room, rendering the flowers with the freshest dewdrops". The young arrivé was invited to join the cardinal's household on the Palazzo Madama.

Caravaggio's inaugural painting of this carafe is lost — but the motif became a set-piece in compositions involving pretty youths. It appears in *The Lutenist*, the restored version of which will be unveiled in Munich next month, and in two versions of *The Boy Bitten by a Lizard*. No one had previously come close to such scintillating and convincing effects.

A window reflected on the left of the vessel is mirrored upside-down inside the right wall, as if in a concave mirror. Cutting across this reflection is a band of superficial glare, accompanied by smaller smudges of brightness. The radiance casts a glimmer across the water, and catches on the shady plant stems. Four speckles of water on the glass perform optical games that mirror those in the carafe. The upper rim of water is bright with captured light, while the foremost stem drags it into a small peak, bearing witness to surface tension.

How did he achieve such effects? We could say he looked hard. But other artists had looked equally hard. To see simultaneously such effects of multiple reflection and refraction is impossible. David Hockney has reasonably suggested that Caravaggio had recourse to lens-



Encore! The carafe of flowers on the left of *The Lutenist* appeared in several of Caravaggio's paintings.

or mirror-based devices (see *Nature* **412**, 860; 2001). Giuseppe della Porta, in the 1589 edition of *Natural Magic*, described an apparatus that combined a lens and concave mirror to achieve "spectacles for my friends, who admired them with great wonder and astonishment. Even though I gave them the explanations of Philosophy and Perspective they didn't want to believe that these were natural things" (see *Nature* **417**, 794; 2002).

A physical reconstruction of the optical set-up, commissioned by the owner and carried out by myself and my colleagues, showed that the carafe was mounted more or less at a level with the centre of the window to the left. In addition, a surprisingly high source of light — probably from an aperture cut in the ceiling or high in the wall — accounts for the flare on the right wall of the vessel. Because the placing of the carafes in the paintings with the youths is inconsistent with this set-up, we can be sure that the carafe became a cut-and-paste motif, replicated to satisfy patrons' demands for something comparable to the lost painting.

We can understand why members of the del Monte circle were so excited by the new naturalism. They were promoting sciences that

depended on new modes of observation, as advocated in Tommaso Campanella's *Philosophy Explicated According to the Senses* (1591). The cardinal, who was to become a staunch supporter of Galileo, was an enthusiastic practitioner of alchemy (in its most experimental mode), and actively supported Ferdinando de' Medici's efforts to establish a glass-making industry in Tuscany. Caravaggio, for his part, was closely associated with a leading figure in the mirror business in Rome.

In this conjoined context of observational science and 'natural magic', Caravaggio's optical naturalism was bound to flourish, establishing a symbiotic relationship with the ideas of those who were forging new ways in instrumental seeing, without his being overly concerned with their mathematical basis.

Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

The Lutenist can be seen at the Stille Welt: Italienische Stilleben aus drei Jahrhunderten in the Kunsthalle de Hypo-Kulturstiftung exhibition in Munich from 6 December 2002 to 23 February 2003.

Mystery of the mutagenic male

Laurence D. Hurst and Hans Ellegren

Old fathers are the source of more genetic mutations in their offspring than either young fathers or mothers of any age. But the apparently most plausible explanation for this effect might not hold.

Mutations are the raw material for evolution, and the cause of genetic disease. But how do they arise? A clue is the finding that fathers are the source of more mutations than mothers. The usual explanation is that the copying of DNA (replication) is error prone^{1–3}, and that women's reproductive cells suffer fewer such errors than men's. The reason is that eggs are the product of relatively few rounds of replication, all of which are finished before birth, but sperm are the result of many more.

Pre-sperm cells go through a few replications before starting the process of becoming sperm. But to replenish the supply, some of the products of these replications go back to the pre-sperm cell bank. As the manufacture of sperm continues throughout a man's life, the number of replication events experienced by the DNA of cells in this cell bank goes up as men age. Consequently, not only should fathers be the source of most mutations but older dads would be expected to contribute more mutations than younger ones^{1,4}. Writing in *Proceedings of the National Academy of Sciences*, however, Irene Tiemann-Boege and colleagues⁵ question whether this 'male age effect' can be accounted for by replication errors alone.

Tiemann-Boege *et al.* investigated a mutation that results in achondroplasia, the most common form of dwarfism. It had been previously established that fathers are the source of all the mutations responsible for this condition⁶, and that the probability of having an affected offspring increases exponentially as a function of the father's age⁴. If we assume an increasing error rate per replication as males age, this pattern is consistent with the replication hypothesis¹.

Using the polymerase chain reaction to identify the mutation in sperm from males of different ages, Tiemann-Boege *et al.* asked whether they could detect an exponential increase in the frequency of sperm containing the mutation. They could not. Up to age 40, the men in their study all had about the same proportion of sperm bearing the mutation. Over age 55 there was also no increase (although the variance was large). Between ages 40 and 55 there was an increase, but overall there was no exponential male age effect. Further, the absolute increase in the mutation rate in young compared with old males was nowhere near that needed to explain the observed rates of achondro-

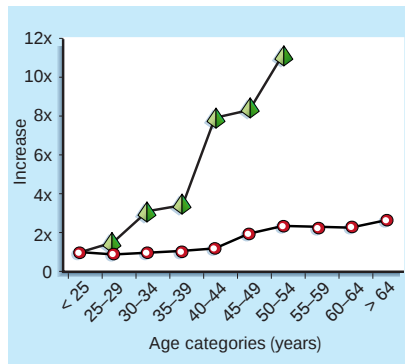


Figure 1 Sperm — not guilty? The increase with paternal age of spontaneous cases of achondroplasia (dwarfism) in offspring is indicated in green. The results of Tiemann-Boege *et al.*⁵, showing the increase — or the comparative lack of increase — with age in the frequency of the achondroplasia mutation in sperm, are in red. On the face of it, the results run counter to the idea that replication errors are one of the main causes of mutation. (Adapted from ref. 5.)

plasia: there was a tenfold difference in the rate of achondroplasia in the children of males aged 20 compared with those aged 50, but only a twofold difference in the frequency of the mutation in sperm (Fig. 1).

What might explain the discrepancy? The answer is not known, but Tiemann-Boege *et al.* propose several alternatives. One is that there is some kind of selective effect. If sperm carrying the mutation are more likely to fertilize the egg, or if fertilized eggs carrying the mutation have a higher chance of success, then the low mutation counts in the sperm could translate to higher rates of affected offspring at birth. Why this might be age-dependent is unclear. Regardless of that, the idea that the exponential age effect owes more to replications, and a higher error rate per replication in older males, is not straightforwardly supported by the findings.

We should, however, hesitate in extrapolating from this one result, not least because the mutation in question may not be typical. For one thing, the DNA at the site of the mutation is modified by the addition of a methyl group, which affects mutational processes. Further, not only is the incidence of the disease-causing mutation remarkably high, but studies of other disease-associated genes indicate that the strength of the male

age effect is highly variable⁴, as is the ratio of male-derived to female-derived mutations³. For example, although all mutations are male-derived in achondroplasia, only 13 of 23 mutations in the gene *NF2* (which result in type 2 neurofibromatosis) are from fathers⁷. This variation is puzzling from the standpoint of the replication model. Perhaps, for all these genetic loci, there is variation in the ratio of mutated sperm to affected offspring but no difference in the underlying male bias to the mutation rate. It would be valuable to know whether, in other diseases in which all mutations are male-derived (Apert's syndrome, for example, which manifests itself as skull, hand and feet malformations), there is the same discrepancy between mutated sperm and affected offspring.

Given the possibility that selection can affect estimates of the rate of occurrence of disease-associated mutations, a better approach to quantifying mutation rates may be to examine 'neutral' mutations — those that have no detrimental effects. This can be done by comparing the same DNA sequence in different species and estimating the rate of neutral evolution. If, in a stretch of DNA, natural selection does not affect the fate of the mutations, then the rate of evolution of the sequence is equal to the mutation rate. From the replication-errors model, it would be expected that the X chromosome should evolve more slowly than the Y chromosome: the Y is always in males whereas, in evolutionary terms, the X spends two-thirds of its time in females. The other chromosomes (the autosomes) should have an intermediate rate.

Differences in the rate of neutral evolution of the three classes of chromosome can then be used to estimate the extent of the sex bias in the mutation rate⁸. If nearly all mutations are replication-dependent, these estimates should accord with anatomically derived estimates of the differences in the number of replication events that precede egg and sperm production. Some calculations have conformed with these expectations, with estimates⁹ that in humans the ratio of the mutation rate in males to that in females (α) is around 5. Comparisons in rodents ($\alpha=2$), cats ($\alpha=4$), birds ($\alpha=4$) and fruitflies ($\alpha=1$) also seem to tie in roughly with expectations from anatomy². In rodents, however, the ratio depends on which chromosomes are compared: a com-

parison of X with Y results in the figure consistent with expectations, but comparing X with autosomes suggests a much higher figure¹⁰.

Although this latter discrepancy is not apparent from human data, analyses of primate and rodent sequences have revealed a further curiosity: not only does mutation rate vary along chromosomes^{11–13}, but also different autosomes have remarkably different rates of evolution^{12,14}. As all autosomes undergo the same number of replications, perhaps this will be the next clue to understanding the causes of mutation. ■

Laurence D. Hurst is in the Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.

e-mail: l.d.hurst@bath.ac.uk

Hans Ellegren is in the Department of Evolutionary Biology, Evolutionary Biology Centre,

Uppsala University, 752-36 Uppsala, Sweden.

e-mail: hans.ellegren@ebc.uu.se

1. Crow, J. F. *Nature Rev. Genet.* **1**, 40–47 (2000).
2. Hurst, L. D. & Ellegren, H. *Trends Genet.* **14**, 446–452 (1998).
3. Chang, B. H. J., Shimmin, L. C., Shyue, S. K., Hewett-Emmett, D. & Li, W.-H. *Proc. Natl Acad. Sci. USA* **91**, 827–831 (1994).
4. Risch, N., Reich, E. W., Wishnick, M. M. & McCarthy, J. G. *Am. J. Hum. Genet.* **41**, 218–248 (1987).
5. Tiemann-Boege, I. et al. *Proc. Natl Acad. Sci. USA* **99**, 14952–14957 (2002).
6. Wilkin, D. J. et al. *Am. J. Hum. Genet.* **63**, 711–716 (1998).
7. Kluwe, L. et al. *Neurogenetics* **3**, 17–24 (2000).
8. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 863–867 (1987).
9. Makova, K. D. & Li, W. H. *Nature* **416**, 624–626 (2002).
10. Smith, N. G. C. & Hurst, L. D. *Genetics* **152**, 661–673 (1999).
11. Matassi, G., Sharp, P. M. & Gautier, C. *Curr. Biol.* **9**, 786–791 (1999).
12. Lercher, M. J., Williams, E. J. B. & Hurst, L. D. *Mol. Biol. Evol.* **18**, 2032–2039 (2001).
13. Smith, N. G. C., Webster, M. T. & Ellegren, H. *Genome Res.* **12**, 1350–1356 (2002).
14. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. *Am. J. Hum. Genet.* **70**, 1490–1497 (2002).

Earth science

Through the wringer

William M. White

Potentially huge amounts of water could be carried deep within the Earth by subducting oceanic crust. But it seems that most of that water is released, fuelling volcanism above subduction zones.

The hydrological cycle usually refers to the process by which water cycles through the atmosphere, rivers and ocean, and round again. But is there another hydrological cycle, one that circulates water through the depths of the Earth? It has long been suspected¹ that surface material can be carried to the Earth's deep interior by subduction of oceanic crust and associated mantle, and then returned to the surface by plumes rising through the mantle. In particular, from isotope data on basalt rocks erupted on oceanic islands, it has seemed that mantle plumes can carry deeply subducted oceanic crust and sediment², as well as the residue of oceanic crust creation (the oceanic lithosphere)³, from depth back to the surface.

Does this process include relatively volatile species, such as water and carbon dioxide? The results reported by Dixon *et al.*⁴ on page 385 of this issue suggest not, at least to any great extent. They find that water is cycled much less efficiently in this way than might be expected — indeed, plumes that apparently contain material recycled from the Earth's surface contain less, not more, water than plumes dominated by the so-called 'common component', such as that beneath Iceland.

Magma readily gives up water at low pressure, the water escaping to the atmosphere during, and even before, eruption. So determining the water content of magma can be difficult. At depths of 1,000 m or more in the

ocean, however, the pressure is sufficient for magma to retain its water as it quenches to glass by contact with sea water. Much of the work on volatiles in magmas has therefore centred on quenched glass recovered from submarine eruptions. These studies have established that when the mantle melts, hydrogen behaves much like the rare-earth element cerium^{5,6}. To get around the variations in water introduced by partial melting and fractional crystallization, the H₂O/Ce ratio can be used to compare the relative amounts of water in different magmas. Dixon *et al.*⁴ determined the water content of submarine glasses erupted on regions of the Mid-Atlantic Ridge thought to be influenced by mantle plumes. The water content of many of these basalts is surprisingly low, the H₂O/Ce ratios being lower than in many other submarine basalts at places such as Easter Island⁷.

Furthermore, samples from the Discovery plume in the South Atlantic and the Great Meteor plume in the North Atlantic display a negative correlation between the ratios of H₂O/Ce and of ⁸⁷Sr/⁸⁶Sr. The significance of this relationship is that it shows that the water deficiency is present in the mantle that produced the basalts, and the strontium isotopes also provide insight into the origin of that mantle. The ⁸⁷Sr/⁸⁶Sr ratio varies in the Earth only because of the very slow decay of ⁸⁷Rb (rubidium) to ⁸⁷Sr. Because the continental crust is enriched in Rb, the crust has a higher ⁸⁷Sr/⁸⁶Sr ratio than

does the mantle, so a high ratio in basalts can signal the presence of recycled crustal material in the mantle. The most water-deficient samples studied by Dixon *et al.* have the highest ⁸⁷Sr/⁸⁶Sr ratios, suggesting that they are derived from mantle containing recycled crustal material.

In this respect, Dixon and colleagues' results are unexpected. One would have thought that recycled material would contain comparatively large amounts of water, for two reasons. First, marine sediments are rich in clay minerals, which can contain 10% or more water bound in their lattices. Second, circulating sea water hydrates the basaltic oceanic crust, eventually raising its structurally bound water content to 2–5% by weight. Additional water is present in the pore space in both sediments and basalt. These water concentrations are far above the ambient concentrations in the mantle, of perhaps 0.03%. The amount of water carried into the mantle in subducted oceanic crust and sediment exceeds 1 × 10¹² kg yr⁻¹ (ref. 8), enough to drain the oceans in little more than a billion years if the water were not returned from the mantle.

Dixon *et al.* argue that the low water content of recycled material results from efficient dehydration of the oceanic crust and sediment during subduction. This idea is not new. Release of water from subducting oceanic crust has long been believed to cause the magma production fuelling volcanoes that ubiquitously sit atop subduction zones^{9,10}. Krakatoa, Mount St Helens, Mount Pinatubo and Soufriere Hills all lie above subduction zones, and the notoriously explosive nature of these and other such volcanoes is largely due to the high water content of their magmas. It is, however, a little surprising that the dehydration process is so efficient. Dixon *et al.* calculate that 92% of the water is extracted from subducting sediment, and 97% from the subducting oceanic crust. The deep hydrological cycle thus appears to be short-circuited, with most subducted water quickly returning to the surface through volcanism rather than being carried into the deep mantle. In other words, the subducted material is effectively put through the wringer before it sinks very far into the mantle.

It would be interesting to compare the flux of water released by subduction-zone volcanoes with estimates of the subduction flux and Dixon and colleagues' calculated dehydration efficiency. It would also be interesting to know if CO₂ is released from subducting oceanic crust and sediment as efficiently as water is. These are difficult tasks, however — determining the water content of magmas erupted on land is problematic, and CO₂ is lost from magma even more readily than water.

Finally, the new results⁴ also point to the importance of subduction-zone processes in

shaping the composition of both the Earth's surface and its interior. Water released by dehydration could carry away much of the soluble-element content of subducting crust and sediment. Any prediction of the composition of deeply recycled crustal material must take account of these losses. ■

William M. White is in the Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York 14853, USA.

e-mail: white@geology.cornell.edu

1. Armstrong, R. L. *Rev. Geophys.* **6**, 175–199 (1968).

2. Hofmann, A. W. & White, W. M. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
3. Schaefer, B. F., Turner, S., Parkinson, I., Rogers, N. & Hawkesworth, C. *Nature* **324**, 304–307 (2002).
4. Dixon, J. E., Leist, L., Langmuir, C. & Schilling, J.-G. *Nature* **420**, 385–389 (2002).
5. Michael, P. *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
6. Dixon, J. E. & Clague, D. A. *J. Petrol.* **42**, 627–654 (2001).
7. Simons, K. & Dixon, J. E. *Geochim. Geophys. Geosyst.* **3**, pap. no. 2001GC000173 (2002).
8. Peacock, S. *Science* **248**, 329–337 (1990).
9. Green, T. H. & Ringwood, A. E. *Contrib. Mineral. Petrol.* **18**, 105–162 (1968).
10. Gill, J. B. *Orogenic Andesites and Plate Tectonics*, 247 (Springer, Berlin, 1981).

Virus evolution

The importance of being erroneous

Sebastian Bonhoeffer and Paul Sniegowski

Viruses must mutate to survive in the face of attack by their host's immune system. A new model suggests that the viral mutation rate is optimized in an evolutionary trade-off between adaptability and genomic integrity.

Evolution by natural selection requires genetic variation, and the ultimate source of this variation is mutation — random errors in genomic replication. In fact, because the fidelity of genomic replication is influenced by genetic variation, the mutation rate is itself subject to natural selection. Broad taxonomic data are consistent with this idea. Although mutation rates per nucleotide vary by factors of up to a million among living things, the variation in genomic mutation rates is less than a factor of a thousand. In RNA viruses, roughly one nucleotide per genome is incorrectly reproduced in each replication; for retroviruses this genomic mutation rate is one per ten replications; and it is one per 300 replications in DNA-based microbes, including DNA viruses and microorganisms¹. Indeed, the remarkable similarity of genomic mutation rates within each of these groups may reflect deep underlying selective constraints, but its explanation remains a challenge to evolutionary biologists.

Writing in the journal *Complexity*, Christel Kamp and colleagues² have now tackled the evolution of genomic mutation rates from a fresh angle. By incorporating the immune response as an explicit selective force in standard 'quasispecies' models, they calculate a viral genomic mutation rate that optimally balances the costs of too much and too little genetic variation.

Quasispecies theory³ was developed over 30 years ago as a means of describing evolution in populations of self-replicating RNA molecules with high mutation rates (quasispecies is the term applied to closely related genetic sequences that are affected as a group by natural selection). But it was soon recognized that quasispecies theory made a useful tool for the study of viral evolution. Its most fundamental prediction is the existence of

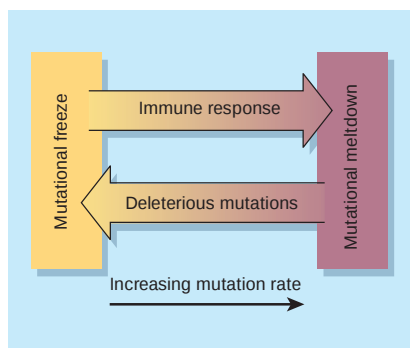


Figure 1 The evolution of viral genomic mutation rates. High mutation rates may enable viruses to escape the host's immune responses, but low mutation rates reduce the probability of destroying essential viral genes. If the mutation rate is too high or too low, the viral population becomes extinct either because the genetic information is irretrievably lost or because the population cannot keep pace with the immune response. Kamp *et al.*² have calculated a genomic mutation rate that optimally balances these constraints.

an error threshold. If the mutation rate exceeds this threshold, then all genomic information is irretrievably lost and the population becomes extinct in a kind of mutational meltdown. In standard quasispecies theory, the simplifying assumption is made that evolutionary fitness is entirely genetically determined and thus constant irrespective of the environment. Under this assumption, a zero mutation rate is optimal and selection should always favour greater fidelity of replication.

But viruses in their natural environments typically face rapidly changing selection pressures as, for example, exerted by the immune response of the body under viral

attack. So Kamp *et al.*² have extended quasispecies theory to incorporate an adaptive immune response. In such an environment, a quasispecies becomes subject to a second mutational threshold, this time a kind of mutational 'freeze': if the mutation rate is too low, then the quasispecies does not keep pace with environmental change and becomes extinct. An optimal genomic mutation rate must therefore lie somewhere between mutational freeze and mutational meltdown (Fig. 1).

Kamp *et al.* have calculated this optimal genomic mutation rate by finding the mutation rate that maximizes viral growth rate in the presence of an immune response. They find that the optimal mutation rate is given by the ratio of the timespan required for the virus to go through an entire replication cycle to the timespan for the immune system to mount a response to a new viral mutant. This result is reminiscent of population-genetic theories that have concluded that a rate of mutation that mirrors the rate of change of the selecting environment is optimal for adaptive evolution^{4–6}.

But Kamp *et al.* go one better than these earlier models in that they incorporate the immune response as an explicit selective force. Comparing their quantitative predictions with data for viral genomic mutation rates, they suggest that many viruses, including HIV, replicate at the optimal genomic mutation rate. Interestingly, their result offers an explanation for the intriguing constancy of genomic mutation rates within viral classes, because the variation in the duration of viral life cycles and the time to mount an immune response is probably considerably smaller than the variation in the rates of mutation per nucleotide.

The idea that viral mutation rates are optimal for escaping host immune responses is appealing⁷, but some questions remain. First, as previously mentioned, genomic mutation rates in RNA viruses are ten times higher than in retroviruses and 300 times higher than in DNA viruses. This doesn't fit the hypothesis of Kamp *et al.*, because there is no clear evidence for systematic differences in the duration of viral life cycles or the dynamics of the immune responses to these classes of virus. Second, escape from the immune system is not a universal feature of viruses: many viruses may survive by transmission to new hosts before the immune response takes effect.

In addition, the model of Kamp *et al.* forgoes some of the realism of population-genetic models for the evolution of mutation rates (reviewed in ref. 8). Such models explicitly consider the fate of genes that modify replication and repair — changes in the frequency of these modifier genes, due to the rise or fall of linked beneficial or deleterious mutations, affect the evolution of the

mutation rate. This indirect selective force is considerably weakened by genetic recombination, a process that breaks apart linked genes — and a factor not included by Kamp *et al.* in their model. But recombination is substantial in many viruses, and its effect should probably be considered explicitly in modelling the evolution of viral mutation rates.

Despite these shortcomings, the paper of Kamp *et al.*² is clearly an important conceptual development in the study of mutation-rate evolution in viruses. Moreover, developing a fuller understanding of the evolutionary causes and consequences of viral mutation rates is worthwhile from both basic and applied perspectives. Drugs that increase genomic mutation rates can kill off viral populations by causing them to exceed their error threshold^{9,10}. A quantitative theory that can predict how close to the error threshold a given viral population is — without the need to estimate its

mutation rate directly — might have real therapeutic value. ■

Sebastian Bonhoeffer is in the Department of Environmental Science, ETH Zürich, CH-8092 Zürich, Switzerland.

e-mail: bonhoeffer@eco.umw.ethz.ch

Paul Sniegowski is in the Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

e-mail: paulsnie@sas.upenn.edu

1. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. *Genetics* **148**, 1667–1686 (1998).
2. Kamp, C., Wilke, C. O., Adami, C. & Bornholdt, S. *Complexity* (in the press); Preprint cond-mat/0209613 (2002), <http://arXiv.org>
3. Eigen, M. *Naturwissenschaften* **58**, 465–523 (1971).
4. Kimura, M. *Genet. Res.* **9**, 23–34 (1967).
5. Leigh, E. G. *Genetics* **73**, 1–18 (1973).
6. Ishii, K., Matsuda, H., Iwasa, Y. & Sasaki, A. *Genetics* **121**, 163–174 (1989).
7. Nowak, M. A. *Nature* **347**, 522 (1990).
8. Sniegowski, P. D., Gerrish, P. J., Johnson, T. & Shaver, A. *BioEssays* **22**, 1057–1066 (2000).
9. Loeb, L. A. *et al. Proc. Natl Acad. Sci. USA* **96**, 1492–1497 (1999).
10. Crotty, S., Cameron, C. E. & Andino, R. *Proc. Natl Acad. Sci. USA* **98**, 6895–6900 (2001).

Plant biology

Fixation with regulation

J. Allan Downie and Martin Parniske

A gene has been isolated that controls the number of symbiotic nitrogen-fixing nodules in legumes. Its similarity to a well-characterized regulatory gene in *Arabidopsis* provides clues about its action.

Leguminous plants produce root nodules, within which symbiotic bacteria capture atmospheric N₂ and convert it into nitrogen that can be used by the plant. But this process is energetically expensive and so legumes strictly control the numbers of nodules they form. Papers by Krusell *et al.*¹ and Nishimura *et al.*² (pages 422 and 426 of this issue), and by Searle *et al.*³ in *Science*, describe the characterization of a regulatory gene that normally limits nodule numbers, and that when mutated increases nodulation. Control of nodule development is of interest in its own right, but this work may also have agricultural applications.

Soybean and pea mutants with enhanced nodulation have been known for about 20 years⁴, but their complex genomes have hampered attempts to clone the genes responsible. Similar mutants^{5,6} were recently identified in *Lotus japonicus*, a legume with a relatively small genome. These mutants have hypernodulation and aberrant roots, hence their designation as *har* mutants. The numbers of both nodules and lateral roots are increased in these *L. japonicus* and soybean mutants, indicating that normal legumes possess a common regulatory mechanism that limits the numbers of root and nodule growing points, or meristems.

Grafting experiments showed that *HAR1* control of nodule and lateral-root number

in *L. japonicus* depends on the shoots rather than the roots (Fig. 1, overleaf), a characteristic that had been previously observed with soybean hypernodulation mutants⁴. So plants with *har1*-mutant shoots grafted onto wild-type roots had increased root nodulation. In contrast, the reciprocal grafted plants (mutant roots with wild-type shoots) had normal roots and nodules. After positioning *HAR1* on a physical map of the *L. japonicus* genome, two groups^{1,2} cloned the gene. They went on to identify mutations in the equivalent genes in pea¹ and soybean² hypernodulation plants, which showed a similar shoot control of nodulation³. Independently, following about 15 years of work³, the equivalent gene controlling nodulation in soybean was isolated and was called *NARK* (nodule autoregulation receptor kinase). It is clear that *HAR1* and *NARK* are the same genes from different species.

HAR1 and *NARK* encode a type of receptor protein that is abundant in plants⁷ and has three components: an extracellular domain of leucine-rich repeats, a membrane-spanning domain, and an intracellular protein kinase domain (Fig. 2). This structure is compatible with the receptor's function being perception of a ligand outside the cell, followed by internal signal transduction through protein phosphorylation by the kinase domain. Mutant genes sequenced from the three species had



100 YEARS AGO

We have received from Messrs. J. W. Gray and Son a pamphlet on scientific protection against lightning, written by Mr. A. Hands. The writer gives a careful explanation of the principles which must be observed in erecting lightning conductors; as the pamphlet is written in non-technical language, it is to be hoped it may be the means of disseminating information amongst the public, since there are few subjects on which more ignorance and superstition exist. The importance of careful protection may be gathered from the fact that Mr. Hands estimates the damage caused annually by lightning in this country alone at from 50,000*l.* to 100,000*l.*

ALSO

The Liverpool correspondent of the *Central News* states that the Nobel prize of 3,000*l.* for researches in connection with malaria will be a personal one to Major Ross, principal of the Liverpool School of Tropical Medicine. According to the Stockholm correspondent of the *Daily Chronicle*, the prize for medicine will be awarded to Prof. Finsen, the Danish discoverer of the treatment by red light for lupus, and the prize for physics to Prof. S. A. Arrhenius. From *Nature* 27 November 1902.

50 YEARS AGO

On November 4 at 16h. 58m. 20s. G.M.T., an earthquake occurred with epicentre... near the east coast of Kamchatka. It was recorded strongly at seismological observatories throughout the world and had a magnitude of 8¹/₃ on the Gutenberg–Richter logarithmic scale... Faulting probably took place in the sea bed near the epicentre since a great tsunami or seismic sea wave resulted, and spread throughout the Pacific Ocean. It arrived at the coast of northern Japan about 20h. G.M.T. on November 4. The Hawaiian warning system was used and the coastlines of several islands, including the Oahu coast, were evacuated in anticipation of the wave. When the wave arrived at Hawaii itself, it is reported to have been several feet high... Waves from one to three feet high arrived at the Whangarei beaches in the north of New Zealand about 7 p.m. local time on November 5. When these waves arrived at Wellington, they were 6–8 in. high... Immediately following the main shock, there were more than a hundred aftershocks. Further news is awaited. From *Nature* 29 November 1952.

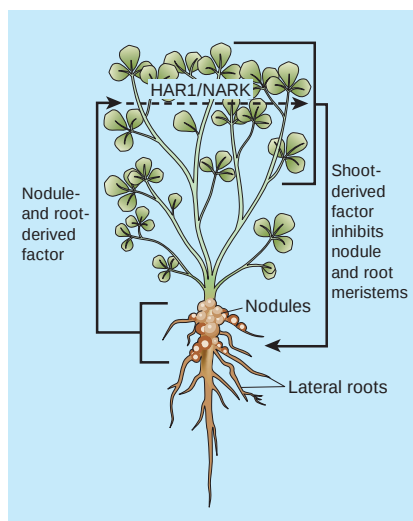


Figure 1 Control of nodule number in legumes. One proposal⁴ is that a factor produced by nodule and root meristems is detected in shoots, resulting in the production of a second factor that suppresses the development of new nodules. In line with this model, mutation of the *HAR1/NARK* gene increases the numbers of nodules and lateral-root meristems^{1–3}. So *HAR1/NARK* protein is required for the production of the shoot-derived inhibitor, possibly through its involvement in perception of a nodule- and root-derived signal.

alterations in the kinase domain, suggesting that phosphorylation by *HAR1/NARK* protein is essential for signalling. The signal perceived by *HAR1/NARK* is unknown, as is the kind of inhibitor that suppresses further nodulation.

Comparison with the model plant *Arabidopsis* provides insight into how *HAR1/NARK* might function, even though *Arabidopsis* does not form nodules. The *Arabidopsis* genome is predicted to contain 216 leucine-rich-repeat, receptor-like kinases⁷. They are believed to perceive extracellular signals such as bacterial flagellin, the wounding signal systemin, and the plant hormones brassinolide and phytosulphokine. A real surprise from the new work^{1–3} is that the legume *HAR1/NARK* is more similar in sequence to *Arabidopsis* *CLAVATA1* than to any other *Arabidopsis* protein, implying that, like *CLAVATA1*, it functions as a receptor kinase. Together with *CLAVATA2*, *CLAVATA1* forms a complex that detects the signal peptide *CLAVATA3* (Fig. 2). The *CLAVATA* genes are involved in regulating cell fate in the shoot apical meristems⁸, *clavata1* mutants having an enlarged meristematic zone that leads to fasciation (contiguous parts growing into one).

CLAVATA1 and *HAR1/NARK* appear to have similar functions, because mutation of the corresponding genes results in increased meristematic activity (Fig. 2), either in shoot apical meristems (*clavata1*) or in root and

nodule meristems (*har1/nark*). However, they must have different roles, because the roots of *clavata1* mutants are unaffected, as are the apical meristems of *nark* mutants³. Moreover, their patterns of expression are different. *CLAVATA1* is exclusively expressed in apical meristems⁸, whereas *HAR/NARK* seem to be expressed in most tissues except apical meristems^{2,3}. It will be interesting to see how far the different functions of *HAR1/NARK* and *CLAVATA1* can be attributed to their expression patterns.

Grafting and other experiments with soybean hypernodulation mutants led to a model of how legumes might regulate nodule number⁴. This postulates that a signal moves from the root to the shoot, and that increased root nodulation is detected in shoots as the root-derived signal increases. As a result, the shoots produce an inhibitor, which is translocated to the root, where it represses further nodule development. Based on the grafting experiments and the nature of *HAR1/NARK*, it seems that this protein could be involved in the perception of a root-made signal in the shoot (Fig. 1). An unexpected result is that all three groups^{1–3} detected *HAR1/NARK* messenger RNA in roots, implying that the protein is being made there. But although the gene is transcribed in roots, the grafting experiments show that the root-expressed gene alone cannot function in nodule repression. So some additional factor in the aerial part of the plant may be required for *HAR1/NARK* to act.

There is a second shoot-controlled hypernodulation gene in pea that, when mutated, can lead to shoot fasciation⁹. Given the similar fasciation seen with the *clavata* mutants, there may be some functional overlap of the *CLAVATA* and *HAR1/NARK* pathways. It seems likely that the product of this second gene in pea could be part of the same signalling pathway as *HAR1/NARK*. Identification of the gene will be required to see if its product can interact directly with *HAR1/NARK* (Fig. 2) and what relationship there is between the *HAR1/NARK* and *CLAVATA* pathways. Genes with high sequence similarity to *CLAVATA1* have been identified in soybean¹⁰, and so *HAR1/NARK* may have arisen as a duplication of an ancestral *CLAVATA1*-like gene, with a subsequent divergence of functions.

Finally, the results of Krusell *et al.*¹, Nishimura *et al.*² and Searle *et al.*³ have practical implications. One of the problems with most of the existing hypernodulation mutants is that the abnormal root and nodule development has severe effects on plant growth. However, Searle *et al.*³ describe one mutation in *NARK* that has less detrimental effects. A different *L. japonicus* gene encoding a transcriptional regulator that represses nodulation has also just been described¹¹, but its relationship to the *HAR1/NARK* receptor pathway is not known. Now that

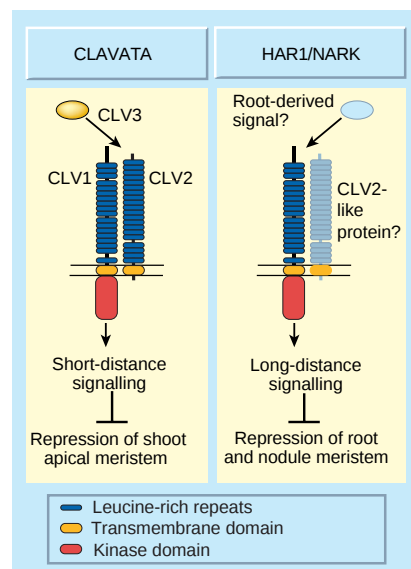


Figure 2 How *HAR1/NARK* might function in signalling and root nodulation. This scheme is based on the model⁸ for signalling by *CLAVATA* proteins in *Arabidopsis*, and on the characteristics of *HAR1/NARK* described by Krusell *et al.*¹, Nishimura *et al.*² and Searle *et al.*³. *HAR1* is most similar to *CLAVATA* (*CLV*) 1, a receptor kinase containing leucine-rich repeats which interacts with *CLV*2, forming a complex that recognizes the signalling peptide *CLV*3. *CLV*2 contains leucine-rich repeat and transmembrane domains, but lacks a kinase domain. *HAR1/NARK* might interact with a *CLV*2-like protein, yet to be identified, to recognize a signal produced by root and nodule meristems.

key genes regulating nodulation have been isolated, it may be possible to identify plants carrying subtle mutations — these might allow increased nodulation, with consequent increases in nitrogen fixation, without affecting plant growth too badly. Such plants might show improved growth, and also leave residual nitrogen in the soil to increase the growth of subsequent crops. ■

J. Allan Downie is in the Department of Molecular Microbiology, and Martin Parniske is in the Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK.

e-mails: allan.downie@bbsrc.ac.uk

martin.parniske@sainsbury-laboratory.ac.uk

- Krusell, L. *et al.* *Nature* **420**, 422–426 (2002).
- Nishimura, R. *et al.* *Nature* **420**, 426–429 (2002).
- Searle, I. R. *et al.* *Science* published online 31 October 2002 (doi: 10.1126/science.1077937).
- Caetano-Anolles, G. & Gresshoff, P. M. *Annu. Rev. Microbiol.* **45**, 345–382 (1991).
- Wopereis, J. *et al.* *Plant J.* **23**, 97–114 (2000).
- Kawaguchi, M. *et al.* *Mol. Plant–Microbe Interact.* **15**, 17–26 (2002).
- Shiu, S.-H. & Bleecker, A. B. *Proc. Natl Acad. Sci. USA* **111**, 10763–10768 (2001).
- Clark, S. E. *Nature Rev. Mol. Cell Biol.* **2**, 276–284 (2001).
- Sagan, M. & Duc, G. *Symbiosis* **20**, 229–245 (1996).
- Yamamoto, E., Karakaya, H. C. & Knap, H. T. *Biochim. Biophys. Acta* **1491**, 333–340 (2000).
- Nishimura, R., Ohmori, M., Fujita, H. & Kawaguchi, M. *Proc. Natl Acad. Sci. USA* **99**, 15206–15210 (2002).

HIV

Bad news for stop–start therapy?

Andrew J. McMichael and Sarah L. Rowland-Jones

An HIV-infected patient who was being treated with anti-retroviral drugs in a 'stop–start' protocol has become infected with a second HIV strain, raising questions about both the treatment strategy and vaccine development.

"This is terrible news," was the response of one senior HIV immunologist at the Barcelona AIDS conference this July, on hearing the results now published on page 434 of this issue by Altfeld and colleagues¹. These authors had studied a patient who was treated with anti-retroviral drugs soon after becoming infected with HIV-1; this therapy achieved good control of his HIV-1 levels. Later, according to the 'supervised treatment interruption' (STI) protocols developed by Walker and colleagues², treatment was deliberately stopped, and then restarted when virus levels rose. This stop–start procedure was repeated twice more, for a total of three treatment-free periods. Altfeld *et al.* found that, during the first two of these periods, the patient's immune system controlled replication of the virus well, if only for a few months. This stands in marked contrast to the situation seen in HIV sufferers who receive anti-retroviral drugs only late in infection; here, within two weeks of stopping therapy, virus levels normally rebound to those seen before treatment.

Viral control during STI is thought to occur through enhancement of the immune response to HIV-1, particularly the response of T lymphocytes that kill HIV-infected cells (these lymphocytes are characterized by expression of the surface protein CD8), and possibly that of 'helper' T lymphocytes (which express the alternative surface protein CD4)³. Indeed, the CD8⁺ T cells from the patient studied by Altfeld *et al.* showed an excellent response, with some 6% recognizing and proliferating in response to HIV-1 proteins. His CD4⁺ T cells also proliferated markedly when exposed to HIV proteins *in vitro* — unusual in HIV-infected people.

So what's the bad news? Things started to go wrong during the third treatment interruption. Virus levels rebounded more rapidly than before, and the response of CD4⁺ T cells deteriorated. On further investigation, Altfeld *et al.* found that a new virus variant had emerged towards the end of the second treatment-free period. Although from the same HIV-1 'family' as the virus causing the original infection — namely the B clade, which predominates in the Western world — this new variant was quite different. The authors carried out a detailed study of 16 regions (epitopes) of HIV-1 that are recognized by CD8⁺ T cells, and found that 7 differed by at least one amino acid between the

original and the new virus; the T cells could not detect these new epitopes. The other 9 epitopes were the same, and the responses of CD8⁺ T cells to them were maintained.

Further analysis of the sequence of the second virus showed that this variant was a distinct strain, probably a new infection (a superinfection) — indeed, the patient reported a recent episode of sexual exposure that was followed by a fever-associated illness. If this is the case, then superinfection took place despite the patient's strong HIV-specific immune response. There is, however, a small chance that the second virus was there from the beginning, possibly hiding in lymph nodes. Although in men the usual pattern is of initial infection with a single virus strain, it is not uncommon for women to be infected with multiple strains from the start — an intriguing example of gender-specific biology in HIV-1 infection³.

But why would a superinfection be such terrible news? It is widely held that CD8⁺ T cells play the major role in controlling HIV replication during a long-lasting infection. This notion is firmly based on evidence such as the rise in viral count that occurs after CD8-blocking antibodies are infused into macaques infected with SIV, the simian version of the virus⁴, and viral mutation to escape CD8⁺ T cells in HIV and SIV infections⁵. These data, together with findings that sex workers who are frequently exposed to HIV seem to be resistant to HIV infection and generate anti-HIV responses through CD8⁺ T cells⁶, have spurred efforts to develop preventive and therapeutic vaccines that stimulate this type of immune response. For instance, vaccinating macaques to induce SIV-specific responses of CD8⁺ T cells enables the animals to control infection more effectively after challenge with an aggressive SIV⁷. So the findings of Altfeld *et al.*¹ might be worrying news for vaccine developers — don't they show that HIV infection can occur in the face of substantial activity of CD8⁺ T cells?

The answer is, not necessarily. First, we do not know the denominator. Recent cases of superinfection have attracted attention⁸, but there may be many more HIV-infected people who can repel an attempted superinfection. Moreover, studies of primates suggest that superinfection is rare: for instance, macaques infected with an attenuated strain of SIV that induces a strong cellular immune response are resistant to

infection with more virulent strains⁹.

Even if the case studied by Altfeld *et al.* turns out to be representative, there are reasons to think that the immune activity generated by a healthy person in response to an HIV vaccine will be qualitatively distinct from that of an HIV-infected person — especially one whose immune system is already damaged by more than three years of infection. For instance, immune impairment can be seen from the very earliest stages of HIV-1 infection, particularly among CD4⁺ T cells; and increases in virus levels after treatment interruption are probably associated with preferential infection of HIV-specific CD4⁺ T cells¹⁰. So the response of CD4⁺ T cells was probably already weakened, before superinfection, in the patient studied by Altfeld *et al.* — indeed, the cells' proliferation when exposed to viral proteins *in vitro* did decline at about the time that the new virus appeared.

CD4⁺ helper T cells may have direct anti-HIV effects, and might also influence the effectiveness of CD8⁺ T cells. Although this patient's CD8⁺ T cells still recognized roughly half of the epitopes in the new virus¹, including one that stimulated the strongest immune response, it seems that the cells failed to protect against superinfection. HIV-specific CD8⁺ T cells in infected people differentiate in an unusual way: they produce little of the membrane-puncturing molecule perforin, and their ability to specifically kill infected cells is consequently reduced¹¹. And the dominant response is not necessarily the most protective, as illustrated by mice infected with lymphocytic choriomeningitis virus¹². Finally, the specificity of the responses of CD8⁺ T cells to SIV and HIV molecules is different in vaccinated macaques¹³ and highly exposed but apparently HIV-resistant sex workers¹⁴ compared with infected animals and people¹⁵. So, the superinfection reported by Altfeld *et al.* might have occurred in the context of a very different immune response from that produced by vaccination of an uninfected person, in terms of T-cell specificity, the competence of CD8⁺ T cells and the strength of CD4⁺ helper T cells.

The new findings¹ may in fact be worse news for the STI strategy than for vaccine development. The usefulness of this strategy very early in infection remains controversial¹⁰. A few patients have achieved remarkable control of HIV-1 and have discontinued therapy entirely, but this seems rare², and as yet there has been no formal double-blind controlled clinical trial — the bedrock of evidence-based medicine. In one study of STI in a long-lasting HIV-1 infection, viral control was not improved despite the enhancement of HIV-specific immune responses¹⁶. The results of Altfeld *et al.* suggest that superinfection with a second HIV strain during a period off therapy could significantly undermine viral control,

so patient commitment to safe sex practices will be an important adjunct to STI. It is not clear whether superinfection is only a risk during treatment interruption: more studies on this are needed.

Altfield *et al.*'s work¹ is a beautiful illustration of the power of modern techniques to explore the minute details of a virus-specific immune response. But the effort required will preclude studies of large numbers of patients. Certainly, this single-patient analysis raises many questions, but whether the news is bad, neutral or even good remains to be seen. Although causing a brief pause for thought, nothing here should slow or divert efforts to develop an HIV vaccine. ■

Andrew J. McMichael and Sarah L. Rowland-Jones are at the MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine,

John Radcliffe Hospital, Oxford OX3 9DS, UK.
e-mails: andrew.mcmichael@clinical-medicine.oxford.ac.uk
sarah.rowland-jones@clinical-medicine.oxford.ac.uk

- Altfield, M. *et al.* *Nature* **420**, 434–439 (2002).
- Rosenberg, E. S. *et al.* *Nature* **407**, 523–526 (2000).
- Long, E. M. *et al.* *Nature Med.* **6**, 71–75 (2000).
- Schmitz, J. E. *et al.* *Science* **283**, 857–860 (1999).
- Klenerman, P., Wu, Y. & Phillips, R. *Curr. Opin. Microbiol.* **5**, 408–413 (2002).
- Rowland-Jones, S. L. *et al.* *J. Clin. Invest.* **102**, 1758–1765 (1998).
- Amara, R. R. *et al.* *Science* **292**, 69–74 (2001).
- Jost, S. *et al.* *New Engl. J. Med.* **347**, 731–736 (2002).
- Daniel, M. D. *et al.* *Science* **258**, 1938–1941 (1992).
- Douek, D. C. *et al.* *Nature* **417**, 95–98 (2002).
- Appay, V. *et al.* *J. Exp. Med.* **192**, 63–75 (2000).
- Gallimore, A. *et al.* *J. Exp. Med.* **187**, 1647–1657 (1998).
- Ferrari, G. *et al.* *Immunol. Lett.* **79**, 37–45 (2001).
- Kaul, R. *et al.* *J. Clin. Invest.* **107**, 1303–1310 (2001).
- Vogel, T. U. *et al.* *J. Immunol.* **169**, 4511–4521 (2002).
- Oxenius, A. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 13747–13752 (2002).

Circadian rhythms

The cancer connection

Michael Rosbash and Joseph S. Takahashi

The *Per2* gene is a core component of the circadian clock in mammals. It now seems that the mouse *Per2* gene is also involved in suppressing tumours, through other genes that affect cell proliferation and death.

If the mantra in real estate is 'location, location, location', in genetics it would be 'phenotype, phenotype, phenotype'. There is simply no substitute for a detailed phenotypic analysis of a mutant strain (study of the overt manifestation of a mutated gene in the organism). This has the potential to reveal unanticipated — and sometimes truly surprising — relationships between genotype and phenotype, or between a primary phenotype and a secondary one. Such was the case for an analysis published recently in *Cell* by Lee and colleagues¹.

The organisms under study here were mice in which both copies of the *mPer2* gene were mutated — a genotype shown previously² to cause a strong defect in circadian rhythms. Most organisms have endogenous 'clocks' that control rhythms of physiology and behaviour with roughly 24-hour (circadian) periodicity. But the period is shortened, and rhythmicity is lost, in the *mPer2* mutant mice. This is reminiscent of the effects of the original *perS* or *perO* mutations in fruitflies, described in the 1971 landmark paper from Konopka and Benzer³.

By observing their mutant mice for a couple of years, however, Lee and colleagues¹ made an unexpected discovery: the animals were unusually cancer prone. At six months of age they began to show excessive cell proliferation in the salivary glands, as well as teratomas — tumours that originate from germ cells and comprise a mix of cell types. Thirty per cent of the mutant mice died before the age of 16 months, half of these

from spontaneous lymphomas. In contrast, such lymphomas were first found in normal mice at the age of 20 months, a highly significant difference. The mutant animals were also more sensitive to γ -radiation, as indicated by premature hair greying and hair loss, and an increased rate of tumour formation — this last effect stemming, at least in part, from a decreased likelihood of cell death (apoptosis) in response to radiation.

So, what could be the story here? The current picture of the central circadian clock in animals is of a self-sustaining transcription–translation feedback loop, involving the transcription of key clock genes, their translation into protein, and the proteins' repression of transcription of the same key genes⁴ (as well as of downstream, clock-controlled genes). In fact, given the importance of post-translational mechanisms — such as protein phosphorylation and turnover — and the lack of translational control in the current picture, it might be more accurate to describe it as a macromolecular feedback loop. In any case, the *mPer2* protein is a key clock component: it contributes to the circadian regulation of transcription of both *mPer2* and downstream genes⁵.

All of which begs the question: is there a tight relationship between γ -radiation and clock genes? And could disruption of circadian transcriptional regulation cause the defects in cell proliferation and death (together termed cell growth) seen even in the absence of γ -radiation in *mPer2* mutant mice? In other words, is there a transcription-

al cascade from clock genes, to downstream growth-control genes, to growth-effector genes? The answers all appear to be yes.

Lee and colleagues' results show that, in normal mice, the expression of several core clock genes was rapidly and potently upregulated in the liver in response to γ -radiation. But in the mutants this response was absent or severely attenuated. Even more surprising was the authors' analysis of a few key genes concerned with cell growth. They found that expression of the *Myc* gene, as judged by levels of its messenger RNA, was circadian in wild-type liver; but in the mutants the expression pattern was modestly shifted and levels of *Myc* mRNA were dramatically increased. Moreover, experiments in cultured cells suggested that *Myc* transcription is directly regulated by the circadian clock. The authors also looked at the expression of *cyclin D1* and *Gadd45a*, two *Myc*-regulated mRNAs, and found that the levels of both fluctuated in a circadian pattern in wild-type livers; in the mutants, both patterns were altered.

So Lee and colleagues propose that the key effect of inactivating *mPer2* is to de-repress *Myc* expression, leading to excessive cell growth and tumour formation. The effect is exacerbated by γ -radiation, which normally upregulates clock genes and thereby presumably leads to *Myc* repression. This fails in the *mPer2* mutants. If these proposals are true, there are some testable predictions. First, overexpressing *Myc* in an otherwise normal genetic background should have the same growth-promoting effects as mutating *mPer2*. Second, and more important, inhibiting *Myc* expression should suppress tumour formation in *mPer2* mutants.

One caveat is that all of these experiments were performed under conditions of 12 hours' light, 12 hours' darkness, so it could be that the mRNA cycles were merely light-driven, not clock-driven. In this context, it is notable that *Myc* and *Gadd45a* were not identified as cycling genes (although *cyclin D1* was) in three out of four microarray studies of liver mRNAs^{6–9}. But the marked effects of the *mPer2* mutation suggest that, at the very least, there is a strong connection between cell growth and the circadian clock. The discrepancy also reinforces the importance of taking microarray data — especially negative data — with a pinch of salt. In our opinion, careful biochemical analyses are more credible.

More generally, the new results¹ have brought into proximity two previously disparate fields of study: circadian rhythms and cell-growth control. One reason why they have hitherto been infrequent bedfellows is that the mammalian circadian rhythm field has historically focused on the brain and, more narrowly, on the suprachiasmatic nucleus — the region of the hypothalamus that is essential for directing cycles of loco-

motor activity. Most adult neurons do not divide yet manifest potent molecular rhythms, suggesting that these rhythms are controlled separately from cell division. A similar conclusion stems from work on rat fibroblast cells, where molecular rhythms persist even if cell division is blocked¹⁰. And classical work in microbes indicates that rapid cell division can be completely uncoupled from the 24-hour timing of the circadian system in the same cells¹¹.

On the other hand, cell division in microbes can be driven by the circadian cycle when the periodicity of division is not far from 24 hours. This suggests that the circadian system is important for proper growth control, and is consistent with the apparent circadian regulation of cell proliferation and apoptosis¹. Moreover, the striking time-dependent response of wild-type mice to γ -radiation¹ reinforces the potential importance of circadian principles in cancer and cancer therapies. It seems that cancer can be a direct consequence of the absence of circadian regulation. Perhaps this applies to other diseases, too. For instance, shift

workers tend to have increased health problems. These could result indirectly from disruption of physiological systems that are under circadian control — but there might also be a direct connection to the clock. ■

Michael Rosbash is at the Howard Hughes Medical Institute, Department of Biology, Brandeis University, Waltham, Massachusetts 02454, USA. e-mail: rosbash@brandeis.edu

Joseph S. Takahashi is at the Howard Hughes Medical Institute, Department of Neurobiology and Physiology, Northwestern University, Evanston, Illinois 60208, USA.

e-mail: j-takahashi@northwestern.edu

1. Fu, L., Pelicano, H., Liu, J., Huang, P. & Lee, C. C. *Cell* **111**, 41–50 (2002).
2. Zheng, B. H. et al. *Nature* **400**, 169–173 (1999).
3. Konopka, R. J. & Benzer, S. *Proc. Natl Acad. Sci. USA* **68**, 2112–2116 (1971).
4. Allada, R., Emery, P., Takahashi, J. S. & Rosbash, M. *Annu. Rev. Neurosci.* **24**, 1091–1119 (2001).
5. Zheng, B. H. et al. *Cell* **105**, 683–694 (2001).
6. Akhtar, R. A. et al. *Curr. Biol.* **12**, 540–550 (2002).
7. Panda, S. et al. *Cell* **109**, 307–320 (2002).
8. Storch, K. F. et al. *Nature* **417**, 78–83 (2002).
9. Ueda, H. R. et al. *Nature* **418**, 534–539 (2002).
10. Balsalobre, A., Damiola, F. & Schibler, U. *Cell* **93**, 929–937 (1998).
11. Johnson, C. H. & Golden, S. S. *Annu. Rev. Microbiol.* **53**, 389–409 (1999).

tion as the highest-energy magnet to its high magnetization (mainly from Fe ions) and to its tetragonal structure and Nd ions, which together produce a high degree of magnetic anisotropy. This anisotropy in turn causes high coercivity.

Is there hope of raising the maximum energy product observed to values approaching the upper bound of 144 MGOe? Despite strenuous efforts, a compound with the necessary properties has eluded researchers. But, about a decade ago, an idea emerged that brought new hope. This is the concept of ‘exchange coupling’ between a hard (high-coercivity) material and a soft (low-coercivity) material with a large magnetization. In a two-phase mixture of such materials, exchange forces between the phases mean that the resulting magnetization and coercivity of the material will be some average of the properties of the two constituent phases⁴. But for the exchange coupling to be effective, the relative sizes of the grains of the two materials must be chosen carefully: Kneller and Hawig⁵ showed that the characteristic dimensions of the soft phase cannot exceed about twice the wall thickness of magnetic domains in the hard phase. Typically, this limits the soft phase to grains of about 10-nm diameter. Similarly, the hard phase must have dimensions of this order or the volume fraction of the soft phase will be rather low, thus limiting the magnetization of the composite.

Skomski and Coey⁴ showed that in an ideal exchange-coupled magnet, consisting of aligned grains of $\text{Sm}_2\text{Fe}_{17}\text{N}_3$ and $\text{Fe}_{65}\text{Co}_{35}$, the theoretical upper bound on the energy product is about 125 MGOe. But in the real world, attempts to fabricate two-phase nanostructures with a high energy product have had only limited success. Bulk magnets produced by subjecting the two-phase material to ultra-fast cooling and then annealing, or by mechanical milling, have not achieved energy products beyond about 20 MGOe (ref. 6). There are several difficulties to be overcome: controlling the material structure at the nanoscale, especially to create uniform grain sizes of about 10 nm; aligning the hard grains sufficiently; and ensuring effective exchange coupling between the two phases in all grains through a homogeneous distribution.

In fact, higher energy products have been obtained in thin-film materials⁷ based on iron–platinum compounds. In these, perpendicular anisotropy (and hence high coercivity) is achieved by first preparing nanoscale Fe–Pt multilayers, and then applying a specific thermal processing technique. Unfortunately, however, this technique is not appropriate for making practical bulk magnets. All in all, the work so far has shown the nanostructuring challenges to be formidable.

But now Zeng et al.¹ have devised a method of chemical synthesis with the

Applied physics

Strong magnets by self-assembly

David J. Sellmyer

Newly developed nanomaterials are proving useful in many fields, but materials that make strong permanent magnets are difficult to devise. Progress has been made using a self-assembled mixture of nanoparticles.

Controlled structuring of materials at the nanoscale can enhance some of their properties and widen their range of applications. Magnetic materials, such as recording media, field sensors and memory devices, are advancing rapidly in terms of their miniaturization, sensitivity and other figures of merit. But progress in producing permanent magnets has been limited by the difficulty of finding new compounds with the necessary properties. On page 395 of this issue, Zeng et al.¹ show how self-assembly of mixtures of magnetic nanoparticles makes it possible to fabricate materials with excellent magnetic properties. The process of chemical synthesis and the resulting nanostructures, with magnetic coupling between nanoscale grains in the composite material, are impressive.

The figure of merit by which permanent-magnet materials are judged is the energy product — a measure of the maximum magnetostatic energy that would be stored in free space between the pole pieces of a magnet made from the material in question². The energy product depends on the area of the ‘hysteresis loop’ (Fig. 1). A typical hysteresis loop arises from plotting the magnetization of the material as the applied

magnetic field is varied — the response of the materials follows two distinct paths on magnetization and demagnetization. As well as the saturation point (maximum magnetization), the hysteresis loop is characterized by the ‘coercivity’ of the material, which is the reverse-field strength needed to reduce the flux density to zero. To obtain a large energy product requires large magnetization and large coercivity.

The largest room-temperature magnetization for a material — a $\text{Fe}_{65}\text{Co}_{35}$ alloy — corresponds to an internal flux density or magnetic induction of about 24 kilogauss. Theoretically, the upper bound on the energy product for the material (proportional to the saturation magnetization squared) is 144 MGOe, for induction measured in gauss (G) and field in oersted (Oe). But in practice, the Fe–Co alloy does not have such a large energy product because it is a soft magnet — that is, its coercivity is actually fairly low and a small reverse field is sufficient to reduce the magnetization to zero.

The largest energy product observed in nature, for the compound $\text{Nd}_2\text{Fe}_{14}\text{B}$, is 56.7 MGOe (ref. 3). $\text{Nd}_2\text{Fe}_{14}\text{B}$ has a complex structure containing 68 atoms in its unit cell. The Nd–Fe–B class of magnets owes its posi-

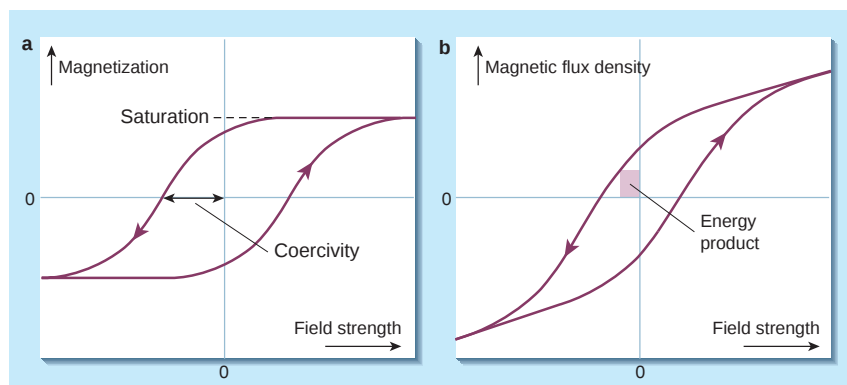


Figure 1 Typical hysteresis loops for a permanent magnet. The response of a magnet to an applied field is shown in terms of its magnetization (a) and the magnetic induction or flux density (b). If the applied field strength is increased until the magnetization saturates, and then reduced, the variation of the magnetization and the flux density traces a hysteresis loop. The coercivity of a magnet is defined as shown in a, and is a factor in the figure of merit of permanent magnets, known as the energy product (shaded area in b).

potential for making three-dimensional magnets with high-energy product. When FePt and Fe₃O₄ nanoparticles of similar sizes (about 4 nm) are mixed and allowed to self-organize, they form structures which, when heated and chemically reduced, form 5-nm-scale homogeneous mixtures of a hard tetragonal FePt phase and a high-magnetization soft Fe₃Pt phase. The admixture of Pt in the Fe₃Pt nanograins results from sintering at a temperature of 650 °C, which is used to induce the phase transition from disordered FePt to ordered tetragonal FePt. The energy product of the two-phase material is 20.1 MGOe—considerably higher than the value expected for isotropic FePt particles alone (13 MGOe).

There are still many practical challenges to be faced if energy products approaching the magic number of 144 MGOe are to be achieved. For instance, ways of compressing the two-phase material into a high-density compact must be explored, as well as

improved alignment of the axes of the hard grains to exploit the full magnetic potential of the nanocomposite system. Nevertheless, the work of Zeng *et al.*¹ is an exciting development that shows the way to making strong magnets for practical applications. ■

David J. Sellmyer is in the Department of Physics and Astronomy, and the Center for Materials Research and Analysis, University of Nebraska, Lincoln, Nebraska 68588-0113, USA.
e-mail: dsellmye@unlnotes.unl.edu

1. Zeng, H., Li, J., Liu, J. P., Wang, Z. L. & Sun, S. *Nature* **420**, 395–398 (2002).
2. Skomski, R. & Coey, J. M. D. *Permanent Magnetism* (Inst. Physics, Bristol, 1999).
3. Rodewald, W. *et al.* in *Rare Earth Magnets and Their Applications* (eds Hadjipanayis, G. C. & Bonder, M. J.) 25–36 (Rinton, Princeton, 2002).
4. Skomski, R. & Coey, J. M. D. *Phys. Rev. B* **48**, 15812–15816 (1993).
5. Kneller, E. F. & Hawig, R. *IEEE Trans. Mag.* **27**, 3588–3600 (1991).
6. Liu, J. P. in *Nanophase and Nanostructured Materials* Vol. 3 (eds Wang, Z. L., Liu, Y. & Zhang, Z.) 230–267 (Tsinghua Univ. Press/Kluwer, 2002).
7. Liu, J. P., Luo, C. P., Liu, Y. & Sellmyer, D. J. *Appl. Phys. Lett.* **72**, 483–485 (1998).

Neuroscience

How neurons compute direction

Peter Sterling

Certain retinal neurons fire specifically in response to stimuli moving in one direction. This apparently occurs when branches of an upstream nerve cell respond asymmetrically, and link asymmetrically to the firing retinal neuron.

Many of the nerve cells associated with vision respond asymmetrically to motion—they fire strongly when an object moves in one direction (the ‘preferred’ direction), but weakly or not at all when the object moves the other way (in the ‘null’ direction). This startling property is displayed by neurons in the cerebral cortex, and by certain neurons in the retina (ganglion cells) that relay visual information

to lower brain centres concerned with eye movements. Because retinal neuronal circuits are simpler and more experimentally accessible than cortical circuits, almost 40 years have been devoted to learning how they compute directional selectivity. Three new papers^{1–3}, including one on page 411 of this issue, seem to have almost solved this maddening puzzle.

Barlow and Levick⁴ were the first to iden-

tify the basic computational process (Fig. 1a, overleaf). Studying the directionally selective ganglion cell in the intact eye of a rabbit, they flashed two slits of light, either alone or separated in space and time, to simulate motion. When motion was simulated in the preferred direction, by the sequence A then B, slit B evoked more firing (the cell was more excited) than when it was presented alone. When motion was simulated in the null direction, by the sequence B then A, slit A evoked less firing than when presented alone (this is inhibition). This led to a simple suggestion: excitation evoked by motion in the preferred direction must reach the ganglion cell before inhibition can cancel it; and inhibition evoked by motion in the null direction must arrive before excitation and cancel it. But how might this occur?

Various clues were gradually pieced together. First, the directionally selective ganglion cell proved to be identifiable by its characteristic branching pattern⁵. Second, these branches were found to closely intertwine with those of a special type of neighbouring neuron, named the starburst cell⁶ (for obvious reasons; see Fig. 1c). And third, the starburst cell was discovered to release the inhibitory neurotransmitter GABA⁷, which is essential for directional selectivity⁸—all of which suggested some role for the starburst cell in computing direction.

Yet these findings provided little clue to directional selectivity, because both starburst and ganglion cells branch out symmetrically from their centres, and a stimulus swept across a starburst neuron evokes the same response in all directions⁹. Furthermore, the heroic laser-induced ablation of most starburst neurons on the ganglion cell’s null side (the side first stimulated by motion in the null direction) failed to affect directional selectivity¹⁰. Thus, evidence seemed to be accumulating against a specific role for starburst cells in computing direction. But when Yoshida *et al.*¹¹ used an immunotoxin to ablate all starburst cells, directional selectivity was abolished. This revived the idea that starburst cells are essential.

Attention then turned to a fine-scale asymmetry of the starburst cell. This neuron receives excitatory inputs all along its branches. But its outputs occur only near the tips⁶ (Fig. 1). This led Borg-Graham and Grzywacz¹² to predict an asymmetry of neurotransmitter release from starburst cells—more transmitter should be released after a stimulus moving outward along a branch than after one moving inward. This would imply that individual branches can serve as distinct computational units. Also, Vaney *et al.*¹³ predicted that starburst branches pointing in opposite directions might connect to ganglion cells with opposite preferred directions. If both predictions were true, directional selectivity could be explained.

To test whether starburst branches

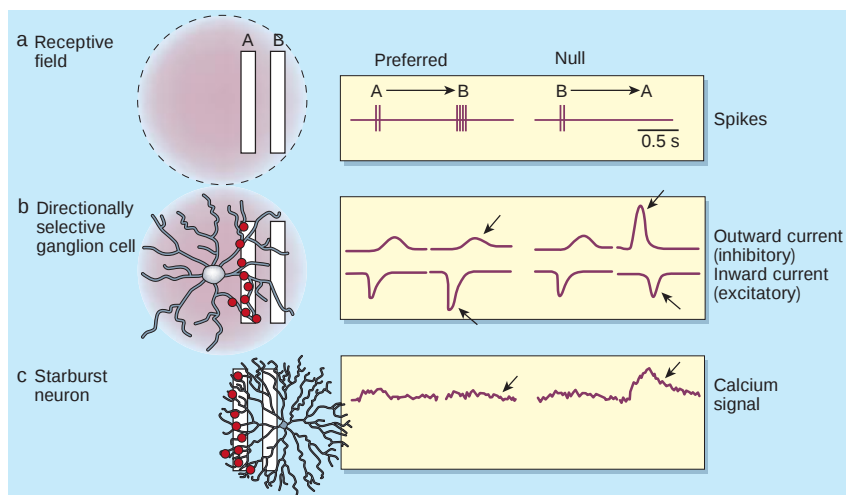


Figure 1 Directional selectivity in the visual system. **a**, Barlow and Levick⁴, recording spikes from a retinal ganglion cell, simulated motion by flashing slits successively. The preferred sequence (A then B) evoked greater firing in response to B. The null sequence (B then A) evoked few or no spikes to A. **b**, Fried *et al.*³ show that, with the preferred sequence, an excitatory input to the ganglion cell starts earlier and rises higher than an inhibitory input. With the null sequence, the inhibitory input starts earlier and rises higher than the excitatory input. The excitatory input is probably suppressed upstream of the ganglion cell. **c**, Euler *et al.*¹ suggest the cause of enhanced inhibition and reduced excitation in response to the null sequence. They show that the preferred sequence, stimulating a starburst cell, evokes no calcium influx at the sites where this cell releases neurotransmitter (red dots). But the null sequence evokes a large calcium influx, which leads to release of inhibitory transmitter (GABA) at these sites. The sites contact branches of the ganglion neuron (red dots in **b**) and thus deliver GABA to that neuron asymmetrically.

operate independently was a difficult task. Electrical currents are normally recorded through a microelectrode attached by suction to the plump cell body, but this integrates currents from all the branches. Yet it is impossible to attach an electrode to just one gossamer-fine branch. Euler and colleagues¹ circumvented this problem by recording responses of single branches optically. Viewing the retina on the stage of a two-photon microscope, they injected a starburst cell body with an indicator dye that diffused throughout the cell. This dye can be excited by coincident, infrared photons to fluoresce in the presence of calcium; thus, whenever calcium levels fluctuate, either spontaneously or after a stimulus, changes in fluorescence can be localized to specific branches.

Euler *et al.* found that spontaneous calcium fluctuations occur in each branch independently of its neighbours. Furthermore, a 'pie-wedge' light stimulus over one sector of the branches evoked a calcium rise strictly in that sector, especially at the output sites. Finally, a stimulus moving outward along a branch did evoke a larger response (in terms of calcium rise) than a stimulus moving inward (Fig. 1c). Because calcium enters via voltage-gated ion channels, its rise in one branch probably reflects depolarization spreading outward along that branch, and the calcium induces neurotransmitter release. Euler *et al.* conclude that this could be the source of directional selectivity — if indeed starburst branches connect asymmetrically to ganglion cells.

Fried *et al.*³ advance both points with new, technically difficult electrical recordings. The technology of the 1960s prevented Barlow and Levick⁴ from observing patterns of excitation and inhibition directly. They could only infer these patterns retrospectively — after excitatory and inhibitory events had been integrated at the ganglion cell's output. But Fried *et al.*³ and Taylor and Vaney² could 'clamp' a cell's voltage at specific levels, and so directly record excitatory and inhibitory currents. They found that the excitatory current is greater in the preferred direction than in the null, and vice versa for the inhibitory current (Fig. 1b). In short, both inputs were themselves directionally selective.

Next, Fried *et al.* showed that slits flashed in the null sequence (B then A) on the null side of the ganglion cell, but well beyond its branches (Fig. 1b), evoke a strong inhibitory current. This would strongly depolarize the tips of starburst branches that reach out towards the ganglion cell, implying that it is their depolarization that causes the inhibitory current. This same sequence and location also reduced the excitatory current evoked by stimuli from the preferred direction, presumably by suppressing release of an excitatory neurotransmitter (presynaptic inhibition). Finally, in a very difficult experiment, Fried *et al.* recorded from a ganglion cell while electrically stimulating a starburst cell whose branches overlapped it. When the starburst cell was on the ganglion cell's null side, an inhibitory current was observed in

the ganglion cell, and the two cells' branches were seen by confocal microscopy to closely entwine. But when the starburst cell was on the ganglion cell's preferred side, electrical stimulation did not affect the ganglion cell, and their branches intersected but did not entwine.

If this last point holds up (only three pairs of each type were studied), the basic mechanisms for directional selectivity would be explained. A stimulus on the ganglion cell's null side and moving in its null direction depolarizes starburst branches that point in that direction. These release GABA onto the ganglion cell and also onto its excitatory inputs. So, as Barlow and Levick⁴ predicted, inhibition would be evoked before excitation reached the ganglion cell branches; and the inhibition would also reduce excitation from the null direction. Similarly, a stimulus on the ganglion cell's preferred side and moving in its preferred direction also depolarizes starburst branches that point in that direction. These starburst branches would also release GABA; however, they do not connect to the ganglion cell in question, but to one with the opposite preferred direction.

All in all, these three papers^{1–3} represent a major intellectual and technical triumph. Yet it would be premature to dismantle the cottage industry that this problem has spawned, because some fascinating questions remain. For example, precisely what biophysical mechanism causes asymmetrical neurotransmitter release by starburst branches? Why is excitation by acetylcholine, another starburst-cell neurotransmitter⁷, not observed in response to the stimuli that trigger GABA release? Finally, by what developmental trick do ganglion and starburst neurons manage to connect asymmetrically? ■

Peter Sterling is in the Laboratory for Retinal Microcircuitry, Department of Neuroscience, 123 Anatomy/Chemistry Building, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6058, USA.

e-mail: peter@retina.anatomy.upenn.edu

1. Euler, T., Detwiler, P. B. & Denk, W. *Nature* **418**, 845–852 (2002).
2. Taylor, W. R. & Vaney, D. I. *J. Neurosci.* **22**, 7712–7720 (2002).
3. Fried, S. I., Münch, T. A. & Werblin, F. S. *Nature* **420**, 411–414 (2002).
4. Barlow, H. B. & Levick, W. R. *J. Physiol. (Lond.)* **178**, 477–504 (1965).
5. Amthor, F. R., Oyster, C. W. & Takahashi, E. S. *Brain Res.* **298**, 187–190 (1984).
6. Famiglietti, E. V. *J. Comp. Neurol.* **309**, 40–70 (1991).
7. O'Malley, D. M., Sandell, J. H. & Masland, R. H. *J. Neurosci.* **12**, 1394–1408 (1992).
8. Caldwell, J. H., Daw, N. W. & Wyatt, H. J. *J. Physiol. (Lond.)* **276**, 277–298 (1978).
9. Peters, B. N. & Masland, R. H. *J. Neurophysiol.* **75**, 469–480 (1996).
10. He, S. & Masland, R. *Nature* **389**, 378–382 (1997).
11. Yoshida, K. *et al.* *Neuron* **30**, 771–780 (2001).
12. Borg-Graham, L. J. & Grzywacz, N. in *Single Neuron Computation* (eds McKenna, T., Davis, J. & Zornetzer, S. F.) 347–375 (Academic, Orlando, Florida, 1991).
13. Vaney, D. I., Collin, S. P. & Young, H. M. in *Neurobiology of the Inner Retina* (eds Weiler, R. & Osborne, N.) 157–168 (Springer, Berlin, 1989).

Rapid and costly ageing in wild male flies

Short-lived insects surprisingly still suffer senescence under natural conditions.

Ageing (senescence) has never been demonstrated convincingly in any insect in the wild, where mean lifespans are probably much shorter than in the laboratory¹, and most evidence for senescence in other wild animals (such as mammals) is limited to their reduced survival with age². Here we show that ageing is detectable in wild populations of a very

short-lived insect, the antler fly (*Protophila litigata*), and causes debilitating and costly effects that force a decline not only in survival probability, but also in the reproductive rate of males. Our findings argue against the possibility of a trade-off between fitness components, whereby survival may decline without senescence if investment in reproduction increases with age³, and indicate that ageing rates are subject to intense selection in the wild.

Although theory predicts the evolution of rapid senescence in organisms that experience high extrinsic (age-independent) mortality rates⁴, it has been suggested that very few individuals in these groups (such as insects or small mammals) survive long enough in the wild to exhibit detectable senescence^{5,6}.

We tested for senescence in a wild population of the antler fly, a small dipteran that breeds exclusively on discarded antlers of moose and deer. The tendency of adult flies to spend their lives on a single antler, as well as the long duration of their mating (2.3 h; ref. 7), facilitate the acquisition of field data on mating success and survival. We surveyed mating aggregations on nine moose antlers every 2 h over 72 days, and recorded the presence and mating status (single or coupled) of each of 609 individually marked males⁸.

The daily probability of mortality in males increased with age (Fig. 1a; likelihood ratio comparisons with constant model: Gompertz, $\chi^2_1 = 4.293$, $P = 0.0383$; two-parameter Weibull, $\chi^2_1 = 3.931$, $P = 0.0474$; three-parameter Weibull, $\chi^2_1 = 3.950$, $P > 0.1$). The Gompertz model fitted marginally better than the two-parameter Weibull, on the basis of Akaike's information criterion (AIC; Gompertz, 78.11; Weibull, 78.29) and bootstrap analysis (Gompertz chosen in 5,838 of 10,000 iterations).

The daily mating rate decreased linearly with age (Fig. 1b; linear AIC, 2,337.94; two-parameter Weibull AIC, 2,338.72; constant AIC, 2,345.03). Although the mean mating rate may decline with age because males that mate less frequently live longer, rather than because of senescence, we found no evidence of a trade-off between mean daily mating rate in early life (age, 2–5 days) and lifespan ($n = 15$ lifespan classes, $r^2 = 0.05$, $F = 0.74$, $P > 0.4$). Mating rate is probably a reliable indicator of male reproductive rate in this species, as take-overs are rare and females lay eggs immediately after copulation⁷. Thus, as predicted by theory⁴, we found that both survival

and reproductive rate declined with age.

To assess the relative fitness costs of senescent declines in survival and reproduction, we compared the reproductive value (that is, the expected lifetime mating success) of males at age 0 days with hypothetical reproductive values based on constant initial (presenescent) daily rates of survival or reproduction. Males lost 1.7 times more fitness as a result of declining reproductive rate than from declining survival rate, but this difference was not significant (bootstrap $P = 0.0992$).

However, a numerical analysis comparing the relative sensitivity of fitness to small, additive reductions in survival and mating rates at different ages of onset⁹ indicated that at younger ages (< 20 days), selection was stronger on survival than on mating rate, whereas at older ages (> 20 days), selection was stronger on mating rate than on survival. Net fitness losses from senescence in both traits amounted to about 20% (95% confidence limit: 13–26%) of potential fitness. However, our analysis may underestimate both total fitness losses and relative losses from senescence in reproductive rate if low-quality individuals are under-represented in older age classes¹⁰ and if male fertility declines with age¹¹.

We have shown that senescence can be detected in wild insects, despite a high extrinsic mortality rate (about 13% per day) and brief median lifespan (6 days). Because there is a simultaneous decline in both survival and reproduction, our results indicate unambiguously that senescence occurs in this wild insect population. The high net fitness costs of senescence confirm a basic assumption of evolutionary theories — that senescence rates are under strong selection in wild animals.

Russell Bonduriansky, Chad E. Brassil

Evolutionary Ecology Group, Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada
e-mail: russell.bonduriansky@utoronto.ca

1. Roach, D. A. *Exp. Gerontol.* **36**, 687–694 (2001).
2. Promislow, D. E. L. *Evolution* **45**, 1869–1887 (1991).
3. Partridge, L. & Barton, N. H. *Proc. R. Soc. Lond. B* **263**, 1365–1371 (1996).
4. Williams, G. C. *Evolution* **11**, 398–411 (1957).
5. Hayflick, L. *Nature* **408**, 267–269 (2000).
6. Kirkwood, T. B. L. & Austad, S. N. *Nature* **408**, 233–238 (2000).
7. Bonduriansky, R. & Brooks, R. J. *Can. Entomol.* **130**, 399–405 (1998).
8. Bonduriansky, R. & Brooks, R. J. *Can. Entomol.* **129**, 827–830 (1997).
9. Abrams, P. A. *Evol. Ecol.* **5**, 343–360 (1991).
10. Service, P. M. *Am. Nat.* **156**, 1–13 (2000).
11. Prowse, N. & Partridge, L. *J. Insect Physiol.* **43**, 501–512 (1997).

Competing financial interests: declared none.

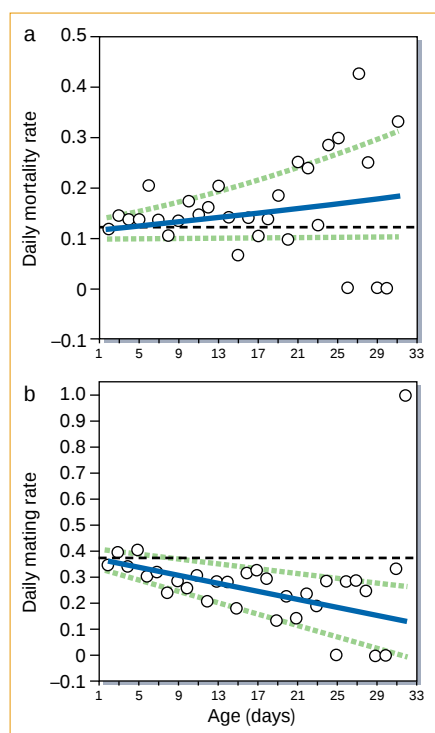


Figure 1 Evidence of senescence in male antler flies (*Protophila litigata*) in the wild. **a**, Mean daily probability of mortality (hazard rate; circles), $h(t)$, with the fitted Gompertz model (solid blue line) and initial hazard rate (dashed black line; instantaneous $h_0 = 0.1359$; $\gamma = 0.0166$). **b**, Mean daily mating success (circles), $m(t)$, with the fitted linear model (solid blue line) and initial mating rate (dashed black line; $\lambda = 0.3741$; $\kappa = -0.0076$, likelihood ratio $P < 0.0001$). Dotted green lines indicate 95% confidence limits. Marked males were released at antlers 24–48 h after adult emergence and were presumed to have died on the day after their last sighting, as mean daily sighting probability was high (63%) and the probability of immigration was very low. Mean lifespan, mating success and mating rate did not change over the season ($P > 0.2$ for all tests). Senescence models were fitted by maximum likelihood, with daily survival and mating rates as random binomial and Poisson processes with age-specific probabilities. Confidence limits were generated by 10,000 bootstraps. Hazard-rate models compared: constant, $h(t) = \lambda$; Gompertz, $h(t) = h_0 \exp(\gamma t)$, fitted as $N_{t+1}/N_t = \exp[-(h_0/\gamma)(\exp(\gamma(t+1)) - \exp(\gamma t))]$; three-parameter Weibull, $h(t) = h_0 + \alpha t^\beta$, fitted as $N_{t+1}/N_t = \exp[-h_0 - (\alpha(\beta+1))((t+1)^{\beta+1} - t^{\beta+1})]$; two-parameter Weibull ($h_0 \equiv 0$). Mating-rate models compared: constant; Weibull; linear, $m(t) = \lambda + \kappa t$. Because data for the mark–release day were incomplete, these data (and the 75 males that survived for less than 1 day after release) were excluded from the analyses.

Thin dielectric films

Uncorrelated breakdown of integrated circuits

In thick dielectrics, electrical breakdown is caused by the generation of spatially and temporally correlated defects that are produced — as in lightning¹ — by feedback between defect formation and local stress². New defects are created in the vicinity of existing defects, leading to rapid and correlated propagation of field-induced defect chains that cause breakdown. By contrast, we show here that defects formed in ultra-thin films subjected to realistic electrical stress remain spatially and temporally uncorrelated, even when multiple ‘shorts’ (chains of defects) begin to bridge the thickness of the films. Besides their relevance to different applications of thin dielectric films, our results have positive implications for the scalability of modern integrated circuits^{3–6}.

To investigate the correlation in defect generation and formation of shorts, we electrically stressed the thin gate dielectrics in about 100 field-effect transistors (FETs) by applying a voltage to the gates and then monitoring the terminal currents as defects accumulated (Fig. 1a).

The discrete and irreversible jumps in the gate current (Fig. 1b) that accompany the formation of individual shorts, as current increases through the low-resistivity paths, enabled us to determine the number of shorts (n) present at a given time, and the time of appearance (t_n) of each short. Likewise, the ratio of the discrete increases in the source and drain currents determines the position (x_n) of these shorts⁷.

To analyse the degree of correlation in our measurements, we used the theoretical framework of a three-dimensional cubic-lattice percolation model^{8,9}. If a film of thickness T_{film} and area A_{film} is stressed at constant voltage, and if the defect generation is spatially and temporally uncorrelated, then the theory makes two key predictions.

First, the cumulative probability distribution of the distances between pairs of shorts should be characterized by the correlation distance $L_c \rightarrow 1$ (M.A.A. and R.K.S., unpublished results). The measured locations of the shorts are characterized by an effective correlation distance of $L_c^{\text{eff}} \geq 0.83$, indicating that the spatial correlation is weak. Second, a single-parameter scaling relation

$$F_n = 1 - e^{-\chi \left(\sum_{k=0}^{n-1} \chi^k / k! \right)} \quad (1)$$

where F_n is the fraction of samples with at least n shorts at time t , must hold. The scaling parameter $\chi = KA_{\text{film}} t^\beta$, where K is the stress-dependent constant, and $\beta \propto T_{\text{film}}$. The results shown in Fig. 1c for each F_n can be inverted by using equation (1) to obtain χ , and then used to predict $F_{1-n} = 1 - e^{-\chi}$.

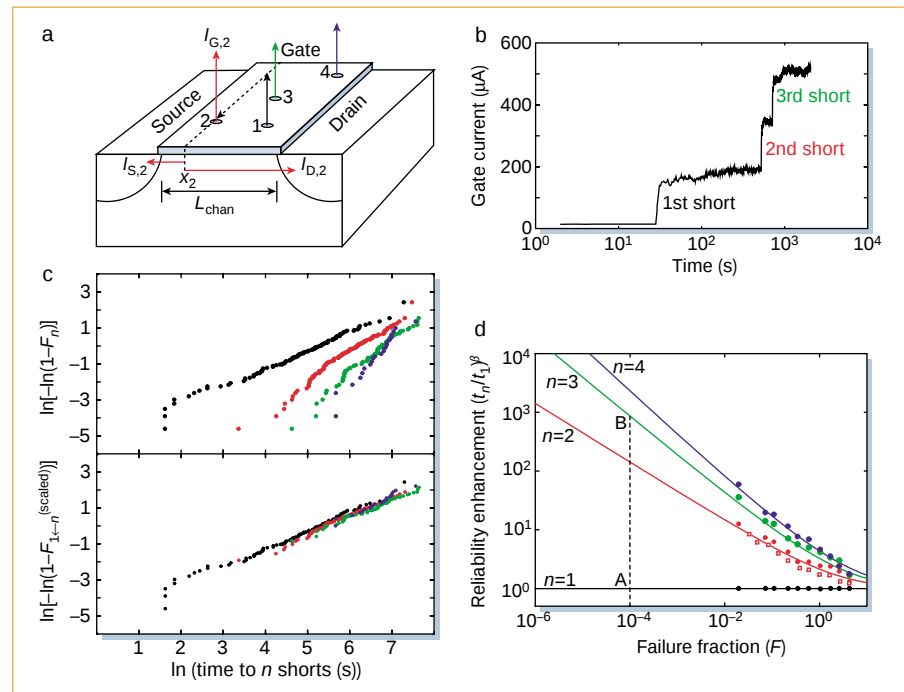


Figure 1 Uncorrelated breakdowns in thin dielectric silica films. **a**, Gate dielectric of a field-effect transistor (shaded) configured to study the evolution of n shorts (n is 1, 2, 3 or 4) and the underlying correlation in defect generation in thin films. Each new short is registered by a discrete jump in the gate current, $I_{g,n}$, balanced by discrete increases in source and drain current, $I_{s,n}$ and $I_{d,n}$, which can be used to time and locate the shorts (for example, short 2 occurred at a distance x_2 from the source along the channel length L_{chan}). **b**, Timing of shorts as a function of gate current. **c**, Measured time sequence for samples with identical area and thickness to develop at least n shorts. Top, traces measured as for **b** are used to compute F_n , the cumulative failure distributions with at least n measured shorts (n is 1–4; $T_{\text{film}} = 1.7$ nm, $A_{\text{film}} = 2.5 \mu\text{m}^2$); bottom, the coincidence of the F_{1-n} , derived from F_n using equation (1), shows that these shorts are uncorrelated in location and time ($n=1$, measured; when n is 2–4, scaled). **d**, Improvement in functional reliability as a function of F . The time, t_n , at which a fraction of integrated circuits with n faults (F_n) fails depends on the ability of the FET to survive $n-1$ shorts without logic failure. Data from oxides of different thickness and area (filled and open symbols), stressed under different conditions, support the scaling theory (solid line). If only 1 out of 10,000 integrated circuit failures is acceptable for a technology, A→B shows the relative improvement in reliability provided that the circuit can sustain just two shorts.

The coincidence of these plots in Fig. 1c shows that equation (1) describes the data well. We find that equation (1) holds for films of different thickness and area, subjected to electrical stresses of different duration and field strength, indicating that the spatial and temporal distributions of the shorts in thin films are essentially uncorrelated.

Historically, silicon FETs used thicker silica films as gate insulators. For these films, the initial short was so catastrophic and subsequent defect generation so strongly correlated that the FET, and consequently the integrated circuit, could no longer perform logic operations. By contrast, modern FETs⁴ use thinner films operated at lower voltages, so that the individual uncorrelated shorts are too ‘soft’ to cause logic malfunctions¹⁰.

If an FET is fault-tolerant up to $(n-1)$ shorts, the operating lifetime (t_n) of the integrated circuit will increase geometrically (from equation (1)) over the traditional fault-intolerant lifetime (t_1); that is, $(t_n/t_1)^\beta \approx (n/e)(2\pi n)^{(1/2n)}/F_n^{(1-1/n)}$ (Fig. 1d). For $\beta \approx 1$, an acceptable failure fraction of $F_n = 10^{-4}$ (point A, Fig. 1d), and an integrated circuit that is fault-tolerant to two shorts, the integrated circuit will function for almost 1,000 times as long as would be predicted on the basis of the model for bulk dielectric

breakdown ($t_3 \approx 850t_1$; point B, Fig. 1d).

Our results indicate that most modern integrated circuits may have been unintentionally overdesigned for reliability, and that it is possible to design faster integrated circuits with higher operating voltage without sacrificing functional reliability. Although issues such as power management, shallow junction formation and so on remain and must be addressed^{5,6}, our findings allow a relaxation of the speed–reliability constraint that effectively removes oxide breakdown as a fundamental obstacle to continued scaling of silicon integrated circuits.

Muhammad A. Alam, R. Kent Smith, Bonnie E. Weir, Paul J. Silverman

Agere Systems, Murray Hill, New Jersey 07076, USA
e-mail: alam@agere.com

- Ogawa, T. in *CRC Handbook on Atmospherics* Vol. 1 (ed. Volland, H.) 23–63 (1982).
- Hippel, A. J. *Appl. Phys.* **8**, 815–832 (1937).
- Moore, G. E. *Electronics* **38**, 114–116 (1965).
- Semiconductor Industry Association Roadmap for Semiconductors 1999 edn (www.itrs.net/itrs/publntrs.nsf).
- Packan, P. A. *Science* **285**, 2079–2080 (1999).
- Kingon, A. I., Maria, J.-P. & Streiffer, S. K. *Nature* **406**, 1032–1038 (2000).
- Degraeve, R., Kaczor, B., De Keersgeiter, A. & Groeseneken, G. *Proc. Int. Reliability Phys. Symp.* **36**, 360–366 (2001).
- Capehart, T. W. & Fisher, M. E. *Phys. Rev. B* **13**, 5021–5038 (1976).
- Stathis, J. J. *Appl. Phys.* **86**, 5757–5766 (1999).
- Alam, M. A., Weir, B. E. & Silverman, P. J. *IEEE Trans. Elec. Dev.* **49**, 239–246 (2002).

Competing financial interests: declared none.

Respiration in the open ocean

Paul A. del Giorgio^{*†} & Carlos M. Duarte^{†‡}

^{*} Département des sciences biologiques, Université du Québec à Montréal, CP 8888, succ Centre Ville, Montréal, Québec H3C 3P8, Canada

[‡] IMEDEA (CSIC-UIB), Instituto Mediterráneo de Estudios Avanzados, C/Miquel Marqués 21, 07190 Esporles (Islas Baleares), Spain

[†] These authors contributed equally to this work

A key question when trying to understand the global carbon cycle is whether the oceans are net sources or sinks of carbon. This will depend on the production of organic matter relative to the decomposition due to biological respiration. Estimates of respiration are available for the top layers, the mesopelagic layer, and the abyssal waters and sediments of various ocean regions. Although the total open ocean respiration is uncertain, it is probably substantially greater than most current estimates of particulate organic matter production. Nevertheless, whether the biota act as a net source or sink of carbon remains an open question.

Oceanic primary production (P) represents about half of the planet's primary production¹, ranging from 35 to 65 gigatonnes of carbon per year (Gt C yr^{-1}) (refs 1–5) with open ocean production accounting for over 80% of the total². Only a small fraction of this organic matter is buried in the ocean's sediments, the bulk of the organic matter produced being instead remineralized through respiration³. Planktonic microbes, particularly heterotrophic bacteria, are responsible for a large fraction of the respiration in the water column of the ocean, particularly in the least productive areas⁶. Although this microbial respiration is supported by dissolved organic carbon (DOC), researchers have traditionally focused on particulate carbon flux⁷, which represents only a fraction of the total carbon flux in the ocean⁸. Approaches based on particle fluxes may therefore underestimate the global ocean carbon flux and specifically, the respiration term that is often derived from these models⁹. Planktonic respiration has been measured directly after changes in O_2 or CO_2 in enclosed water samples^{6,10}, but this approach is effective only in areas where metabolic rates are higher, such as in surface waters. As a result, respiration (R) remains the least constrained term in most models of metabolism, gas exchange and carbon mass balance in the ocean⁴.

A major assumption in contemporary oceanography has been that, over a year and across large spatial scales, primary production must be nearly balanced by planktonic respiration. But direct estimates of microplankton oxygen consumption in oceanic surface waters have challenged this assumption^{10–13} and several recent comparative studies have suggested that respiration may systematically exceed production in large areas of the oceans^{14,15}. The ensuing debate^{16,17} has focused on the processes that occur in the uppermost, illuminated layer only, which covers less than 6% of the ocean volume. Mass balance and carbon export studies generally suggest a modest contribution of intermediate waters to total R relative to that of surface waters⁷. However, direct metabolic measurements systematically suggest that a large fraction of the total carbon mineralization may occur in intermediate waters^{18,19}.

Here, we review the current information on the contribution of various biotic components and depth layers to respiration in the open ocean. We examine estimates of microplankton respiration in the ocean surface, which comprises the illuminated or photic zone (where primary production takes place), the mesopelagic layer, and the ocean interior ($>1,000$ m), which includes the respiration of benthic communities in the ocean floor. We also assess the possible contribution of metazoan zooplankton and larger metazoans to respiration in the water column. Finally, we evaluate the probable magnitude of total respiration, as well as the extent of the remaining uncertainties.

The photic layer

The photic layer extends roughly to the depth receiving 1% of the surface illumination, located at less than 200 m depth in most oceanic areas. The respiration rate in the surface layer is scaled to

the primary production^{11–14}. Some results^{14,15} suggest a tendency for respiration to exceed production in the photic layer of oligotrophic oceans, whereas some others¹⁶ suggest an overall balance between production and respiration in the photic layer of the different regions of the ocean. Although there is an overall positive relationship between respiration and the level of system productivity over large spatial scales, there is often a large degree of uncoupling between the two processes at smaller scales^{10,20,21}. Rates of respiration in the surface layer of the ocean are typically high, with average values of about $1.2 \text{ g C m}^{-2} \text{ d}^{-1}$ (refs 15, 16), which, if extrapolated over the $290 \times 10^6 \text{ km}^2$ of the open ocean, represent a global respiration of about 143 Gt C yr^{-1} . This estimate is about three to four times the accepted estimates of production ($35\text{--}65 \text{ Gt C yr}^{-1}$) (refs 1–5) because the available estimates of respiration in the photic layer are biased towards more-productive regions and periods of high metabolic activity¹⁷.

Intermediate waters

The mesopelagic layer extends roughly from below the base of either the photic layer or the mixed surface layer, generally below 150–200 m, to a depth of approximately 1,000 m, and contains the thermocline where steep chemical and physical gradients occur. Although early studies²² found measurable respiration rates below 200 m in the open Atlantic and Pacific, the mesopelagic and abyssal respiration was traditionally regarded as insignificant²³. But the intermediate waters within the thermocline are now believed to be the main site for organic matter mineralization, which is necessary to re-supply the photic layer with inorganic nutrients²⁴.

The volumetric rates of respiration decrease sharply below the base of the photic layer, and often remain low throughout the thermocline, typically at 1–30% of the rates found in the overlying photic layer^{18,22}. Yet, the depth-integrated microplankton respiration over the much thicker layer covered by the mesopelagic waters often reaches values comparable to those in the photic layer. Models predict^{25,26} that the integrated respiration rate in the 200–1,000 m layer should range from 30 to 130% of the integrated respiration in the 0–200 m layer. These predictions are broadly in agreement with *in situ* determinations of respiration in the surface and mesopelagic layers of various oceanic regions^{9,18,24,27–29}.

In addition to relatively high integrated rates, discrete subsurface metabolic maxima, with rates comparable or even higher than those in surface waters, have often been reported within the ocean's thermocline^{19,24,29,30}. In general terms, these deep layers of high metabolism are relatively thin and limited to tens of metres, but they can nevertheless contribute heavily to the total respiration in intermediate waters³⁰.

The ocean interior

The rates of water oxygen consumption below 1,000 m are often extremely low and difficult to measure^{6,22} resulting in rather

few direct measurements of respiration in the ocean's interior. Fiadeiro and Craig³¹ calculated a rate of respiration below 1,000 m of $2 \mu\text{l O}_2 \text{ l}^{-1} \text{ yr}^{-1}$, similar to other estimates available for the ocean interior^{19,32}. These estimates of deep water respiration agree with recent estimates of DOC consumption rates in the ocean interior³³. The combination of estimates of deep-ocean benthic respiration together with estimates of organic matter flux in the water column³⁴ yielded an estimated total respiration in the ocean interior of $1.2 \times 10^{14} \text{ mol O}_2 \text{ yr}^{-1}$, of which benthic respiration represents 45%, in agreement with other independent estimates of deep-ocean respiration^{31–33}.

Metazooplankton and vertebrates

Estimates of the contribution of metazooplankton to respiration in the ocean are highly variable, varying from less than 1% to more than 50% of respiration depending on the oceanic region, productivity, depth, and method of estimation^{22,28,35–38}. Estimates of the contribution of vertebrates to respiration in the open ocean are very few. But assuming that vertebrates occur at least three trophic levels above primary producers and assuming also an efficiency of 10% at each trophic level, vertebrate respiration must be, on average, $\leq 1\%$ of the respiration in the open ocean.

Temporal and spatial variability in open-ocean respiration

The spatial and temporal variation in water-column respiration is still insufficiently characterized in oceanic systems, and this limitation adds further uncertainty to any global or regional estimate. Although there is an overall positive relationship between photic zone respiration and primary production over large spatial scales, respiration appears to be much less variable than other biological processes, including primary production^{10,11,13–15,20,21,39}. One of the reasons ocean respiration is less variable than production is that heterotrophic microorganisms utilize a diversity of organic matter, and not just that derived from contemporary production. Much of the respiratory carbon demand in the open ocean is supported by DOC^{8,9,40}, but this does not necessarily imply exclusive use of recent primary production, because organic matter may persist for months, or hundreds or thousands of years in the dissolved form. The consumption of past primary production, stored in the form of

the very large^{30,41} (about 685 to 700 Gt C) oceanic DOC pool, potentially allows an uncoupling between primary production and respiration extending over millennia, the mean age of the ocean's DOC pool⁴², resulting in a lack of steady-state at shorter timescales.

The oceanic *P/R* ratio may have varied considerably over geological times. For example, there is evidence that in the period between the Last Glacial Maximum and the Holocene epoch the oceans acted as a source of CO_2 , of at least 170 Gt C, to up to 1,350 Gt C (ref. 43), although it is unclear whether this was the result of increased respiration or decreased primary production. Indeed, respiratory processes are closely dependent on temperature⁴⁴, so that variability in water temperature may account for spatial and temporal variability in open-ocean respiration beyond that induced by changes in primary production.

Total respiration in the open ocean

Estimating the total oceanic respiration is important for our understanding of the global carbon cycle. It should also help us to test the consistency of present estimates of global primary production and organic inputs to the different compartments of the open ocean. We have used the information available to provide a first, necessarily imprecise, approximation of the magnitude of the total respiration in the open ocean that may reduce the present uncertainty.

We do not as yet have a representative mean value of respiration for the surface layer of the ocean¹⁷. However, two equations scaling respiration and photosynthetic rates are available, a linear equation¹⁶, and a nonlinear one, which suggests that the *P/R* ratio is lowest in oligotrophic regions of the ocean and increases with increasing primary production¹⁵. We used these two equations to derive a first-order estimate of the total respiration in the photic layer for the open ocean. We applied each of these two equations to the estimates of the average primary production of each of the biogeochemical provinces in the open ocean², and then scaled the calculated average rates of photic layer respiration to the area of each province. We then aggregated the respiration estimates for each province to yield estimates of the total respiration in the surface layer of the open ocean between 32 Gt C yr^{-1} (using the linear equation¹⁶) and 42 Gt C yr^{-1} (using the nonlinear equation¹⁵).

Table 1 Open ocean respiration and organic matter inputs

Component	Low estimate (Gt C yr ⁻¹)	High estimate (Gt C yr ⁻¹)	Central estimate (Gt C yr ⁻¹)
Respiration			
Photic zone*	32	42	37
Mesopelagic†	21	28	24.5
Ocean interior‡	1.3	1.6	1.5
Mesozooplankton§	1.5	4.5	3
Vertebrates			0.01
Total respiration	55.8	76.1	66
Organic matter inputs			
Measured primary production (¹⁴ C-based)	28	52	40
Unmeasured gross production	13.4	25	19.2
Total production	41.4	77	59.2
Import from coastal areas	6	6	6
Atmospheric inputs	3	3	3
Ancient organic matter	0.5	0.5	0.5
Total inputs	50.9	86.5	68.7
Export and new production			
Suess ⁷			<6
Emerson <i>et al.</i> ⁶⁵			<11
Sambrotto <i>et al.</i> ⁶⁶			15
Falkowski <i>et al.</i> ⁶⁷			16
This paper[†]	23	31.8	27.5

*The range results from applying the two available equations (in refs 15 and 16) to scale *R* to *P*.

†Assumed to be two-thirds of photic layer *R* and a minimum error of 35%.

‡A minimum error of 10%.

§Assumed to be on average 10% of total *R* with a minimum error of 50% around central value.

||Assuming 35% algal respiration and 10% DOC release of ¹⁴C-based production.

† Calculated as $R_{\text{mesopelagic}} + R_{\text{interior}} + (0.5 \times R_{\text{zooplankton}})$.

These estimates of photic layer respiration (R_{photic}) suggest a central value of about 37 Gt C yr^{-1} (Table 1) with an uncertainty of at least 15% (that is, $\pm 5 \text{ Gt C yr}^{-1}$).

Most studies show that integrated respiration in the mesopelagic layer is at least of a similar magnitude to the integrated microplankton respiration in the photic layer of the open ocean, and often greater. Based on the published information reviewed above, the integrated respiration in the mesopelagic layer ($R_{\text{mesopelagic}}$) can be conservatively taken to amount to two-thirds of the average microplankton respiration in the photic layer (R_{photic}) of the open ocean. R_{photic} is estimated to be $32\text{--}42 \text{ Gt C yr}^{-1}$, so the total $R_{\text{mesopelagic}}$ for the open ocean is likely to be of the order of $21\text{--}28 \text{ Gt C yr}^{-1}$, with a probable central value of around $24.5 \text{ Gt C yr}^{-1}$ (Table 1). The uncertainty about this best estimate is extremely large: it is affected by the 15% uncertainty about R_{photic} , and the variability about the ratio of $R_{\text{mesopelagic}}$ to R_{photic} . The latter is unknown, but it is unlikely to be less than 20%, thereby resulting in a minimum uncertainty of 35% about the best estimate of $R_{\text{mesopelagic}}$ derived with the information available.

Converted to carbon consumption assuming a respiratory quotient of 1, the estimate of respiration in open ocean waters below 1,000 m derived by Jahnke³⁴ yields an estimate of total microplankton R_{interior} of around 1.5 Gt C yr^{-1} (Table 1). The calculated net rate of microbial utilization of DOC in the deep ocean³³ ($0.5 \mu\text{M C yr}^{-1}$) supports a respiration rate, if extrapolated to the ocean interior, of about 0.7 Gt C yr^{-1} , which is about half of the total respiration in the ocean interior calculated by Jahnke³⁴ and suggests once again that a large fraction of respiration is supported by DOC. The respiration rate within the ocean interior (water column + ocean floor) appears to be rather well constrained, with a 10% variability about the different estimates³⁴.

As with the respiration in the interior of the thermocline, it is difficult to estimate the contribution of metazooplankton to total plankton respiration, but a contribution of 5 to 10% seems most plausible from the information available. We will conservatively assume that $R_{\text{zooplankton}}$ represents 5% of the combined microplankton respiration in the photic and thermocline waters, with total $R_{\text{zooplankton}}$ representing, therefore, about 3 Gt C yr^{-1} (Table 1). A recent meta-analysis of published data⁴⁵ has concluded that approximately 5.5 Gt of phytoplankton carbon are consumed per year in the global ocean, with a respiratory loss of approximately 1.5 Gt C yr^{-1} , but this independent estimate of mesozooplankton respiration is conservative because it does not include the respiration of the non-phytoplankton carbon ingested, which may support a large portion of the zooplankton C demand in open-ocean systems⁴⁵.

The information on vertebrate respiration in the open ocean is even sparser, and only indirect calculations can be used to derive a global view of its importance. The total reported fisheries catches amount to approximately $0.3 \times 10^{12} \text{ g C yr}^{-1}$ for the oceanic gyre systems⁴⁶. The total fish production can be approximated assuming that the catch represents 20% of the total fish production⁴⁶, which would then amount to $0.0015 \text{ Gt C yr}^{-1}$. Assuming fish respiration to be 9-fold greater than their production, an estimate of only $0.0135 \text{ Gt C yr}^{-1}$ is derived, far below the estimated planktonic respiration in the open ocean and of little significance at a global scale.

The sum of the components discussed above results in a likely range for respiration in the open ocean of 56 Gt C yr^{-1} to 76 Gt C yr^{-1} , with a central value for total respiration of about 66 Gt C yr^{-1} (Table 1). This estimated range of open ocean respiration represents a crude first-order approximation, because in addition to the paucity of data, there are many problems associated with the direct measurement of respiration in open-ocean waters^{6,10,40}, and with the extrapolation of these measurements to larger scales⁴⁷. But these estimates are also conservative, thereby defining a probable lower range of values, and offer a basis for

comparison with other better understood aspects of ocean metabolism.

The metabolic balance

Current estimates of primary production in the oceans vary widely, but most estimates for the global ocean are in the range of 35 to over 65 Gt C yr^{-1} (refs 1–5). Open-ocean production accounts for at least 80% of the total², so most current estimates place open-ocean productivity in the range of $28\text{--}52 \text{ Gt C yr}^{-1}$ (Table 1). The likely range in total open-ocean respiration of 56 Gt C yr^{-1} to 76 Gt C yr^{-1} is higher than the accepted range in global oceanic production, but this apparent imbalance does not imply that either range of estimates is wrong. The validity of global estimates of planktonic primary productivity obtained from satellite remote sensing depends greatly on the quality of calibration data, which are based on ^{14}C incorporation into particulate material⁴⁵. There are two major problems associated with ^{14}C -based production estimates that may result in an underestimation of primary production and in the apparent discrepancy with oceanic respiration. First, ^{14}C -based production estimates do not necessarily represent gross primary production (GPP), because some algal respiration has already been accounted for in the estimate. The fraction of GPP lost to algal respiration varies widely but averages at about 35% (refs 48, 49).

Second, the incorporation of ^{14}C into particulate material does not capture the production of new dissolved organic matter by phytoplankton, which can be significant, particularly in unproductive regions^{8,9,50}. The proportion of primary production that is directly excreted as DOC varies widely, and at least part of this variation may be due to uncertainties associated to the actual measurements⁵⁰. Meta-analysis of published data has suggested that DOC production represents, on average, 13% of the ^{14}C incorporation⁵¹ after respiratory losses, and we will use this conservative estimate to correct current productivity estimates for comparison with respiration.

The oceanic primary production estimates based on ^{14}C incorporation measurements can thus be incremented by about 48% to account for the average algal respiration and DOC production (Table 1). This correction yields a revised estimate of the oceanic GPP in the order of 41 to 77 Gt C yr^{-1} (Table 1), comparable to the estimated range of total respiration in the open ocean of 55 to 76 Gt C yr^{-1} derived above. This very rough correction of open-ocean production estimates narrows the gap between oceanic respiration and production, but some discrepancies in open-ocean metabolism still persist.

The surface waters of the vast, unproductive open-ocean gyres appear to be characterized by extremely high respiration relative to primary production, with P/R ratios often ≤ 1 (refs 14, 15). This pattern cannot be explained solely on the basis of underestimation of production by ^{14}C -based techniques, because the same results have been obtained using O_2 -based measurements, which do not underestimate GPP^{10–13,20,21}. The validity of the high values of respiration in the ocean is supported by consideration that bacterial production represents about 25 to 35% of primary production in the photic layer of the open ocean^{3,52}. Given that bacterial carbon growth efficiency in the open ocean tends to be below 30% (ref. 40), bacterial respiration alone in the photic layer must be comparable to, if not greater than, primary production in the open ocean⁴⁰. In addition, depth profiles show that integrated bacterial production in various open ocean regions often exceeds the estimated particulate carbon export from the photic zone^{3,20}. Recent large-scale studies of bacterial metabolism and primary production in the southern Atlantic have shown that large areas in the oceanic gyres are indeed dominated by heterotrophic bacterial metabolism, whereas at higher latitudes the balance is autotrophic⁵³. The apparent imbalance in the open oceans may be simply the result of temporal lags between P and R , or the lateral transport of recently

produced organic matter^{14,16,20}, but in some areas the imbalance is sufficiently persistent to suggest that alternative sources of organic matter must be used. The crucial question is then: what are the sources of organic matter that fuel these relatively high rates of respiration in the surface waters of oligotrophic oceans?

Allochthonous organic inputs to the open ocean

Coastal areas typically produce an excess organic carbon⁵⁴ that, together with organic inputs from land⁵⁵ can be exported to the open ocean and oxidized there by microorganisms. For example, inputs of particulate organic carbon (POC) and, particularly, DOC from ocean margins to the ocean interior may be an order of magnitude greater than inputs of recently produced organic matter from surface waters^{55,56}. It is believed that only a small fraction of the primary production on land (<1%, calculated from the estimated riverine input of about 0.45 Gt C yr⁻¹ and the terrestrial net primary production¹ of about 56.4 Gt C yr⁻¹) reaches the coastal ocean. A rough estimate of the organic carbon available to be transferred from the coastal to the open sea can be derived as the sum of the excess production by coastal ecosystems, estimated to range between 2.7 Gt C yr⁻¹ (ref. 54) and 6 Gt C yr⁻¹ (refs 16, 48), and the difference between the riverine input organic carbon, of about 0.45 Gt C yr⁻¹ (ref. 57) and the 0.1 Gt C yr⁻¹ buried in coastal sediments⁴¹. These approximations result in an estimated organic carbon transfer from the coastal to the open ocean of at most 6 Gt C yr⁻¹. Regrettably, present models of the organic carbon cycle ignore coastal inputs, because the boundary conditions of general circulation models are set at the shelf break, assuming, therefore, no exchange between open and coastal waters.

The total atmospheric organic carbon input to the open ocean is at present unknown, although there are some relatively high estimates⁵⁸. A very rough, order-of-magnitude estimate of the atmospheric organic carbon input can be derived from consideration of the input of dissolved organic nitrogen (DON), estimated to be about $(2-6) \times 10^{12}$ mol N yr⁻¹ (ref. 59), and a conservative atmospheric DOC/DON ratio of about 20, yielding an estimate of atmospheric DOC deposition of 0.48 to 1.44 Gt C yr⁻¹. The C:N ratio of atmospheric dissolved organic matter is likely to be higher, but even the assumption of a C:N ratio of 40 yields a total atmospheric DOC input of up to 3 Gt C yr⁻¹. These values represent minimum estimates, because they do not include POC deposition.

Organic carbon surplus from past episodes of enhanced ocean upwelling and productivity, such as that apparently occurring during the last glacial period⁶⁰ may have entered a long-term dissolved reservoir. A connection between ancient primary production and contemporary respiration could be a plausible mechanism to account for an excess respiration in the open ocean. But this would require a long-term trend for a net decline of the oceanic organic carbon inventory, for which there is no evidence, and the actual estimates of the net consumption of deep-ocean DOC³³ (0.5 μ M C yr⁻¹) are low. Nevertheless, the slow but steady consumption of ancient, recalcitrant organic matter may be an important mechanism buffering fluctuations in open-ocean metabolism by uncoupling respiration from contemporary primary production. The DOC below the photic layer of the ocean is old, typically having originated thousands of years ago⁴². Even though ancient DOC is believed to be refractory, it may be photodegraded to forms more readily used by bacteria upon exposure to intense light when deep oceanic water is returned to the ocean's surface⁶¹. Indeed, radio-carbon evidence indicates that surface water bacteria consume significant amounts of ancient DOC in the open oceans, but less so in the more productive coastal areas⁶². The mobilization of the excess organic carbon buried in the continental margins due to erosion by rising sea level⁶³ during interglacial periods may be another significant source of ancient organic matter fueling present-day respiration in the open ocean that should be further investigated.

Balancing act

Terrestrial, coastal and atmospheric sources may thus account for organic matter inputs to the open ocean of approximately 9 Gt C yr⁻¹, with an additional 0.5 Gt C yr⁻¹ from the use of ancient DOC. Assuming that most of these inputs are oxidized in the water column, they would be sufficient to sustain a small but systematic imbalance between *P* and *R* in the surface waters of the open oceans, particularly where *P* is low.

The total organic inputs to the open ocean can be estimated from the sum of the corrected open-ocean primary production (41 to 77 Gt C yr⁻¹), and the additional allochthonous organic matter inputs (9–10 Gt C yr⁻¹), and probably range from 51 to 86 Gt C yr⁻¹ (Table 1). The total organic inputs to the ocean calculated in this way compare well with the estimated range in open-ocean respiration, which we hypothesize to vary around a probable central value of 66 Gt C yr⁻¹ (Table 1).

Although there is a correspondence between the plausible organic matter inputs and respiration at a global scale, there are incongruities with specific aspects of this budget. The respiration in the dark layers of the ocean represents up to half of the total respiration in the upper layers, and requires, therefore, a significant input of organic matter. The estimates of export production from the photic zone, based on particle flux and mass balances, have systematically increased during the past decade^{7,64–67}. But even the highest current estimates⁶⁷ would still represent approximately half of the apparent organic matter mineralization calculated from direct respiration measurements in intermediate and abyssal waters of the open oceans (Table 1). There is, therefore, a need to revise both the estimates of organic carbon export and mesopelagic respiration to reconcile these estimates.

We suggest that production and export estimates for the open ocean may have to be revised upwards by at least 50%. The traditional emphasis on the flux of POC has neglected the direct production of DOC by phytoplankton, and this component of oceanic productivity is still largely unconstrained. In fact, a significant portion of the carbon flux both within and out of the photic zone may be through carbon-rich DOC^{9,24,50,66}. A priority in the future should be the accurate estimation of the production and export of oceanic DOC.

Oceanic respiration and the global C budget

Oceanic respiration is estimated here at about 55 to 76 Gt C yr⁻¹, but this estimate does not include the respiration in shallow coastal and continental shelf areas, which can be significant⁵⁴. It is clear that oceanic respiration represents a major source of CO₂ in the biosphere, comparable to the estimated 70 to 80 Gt C yr⁻¹ contributed by soil respiration⁶⁸. Soil respiration has been generally regarded as the largest biological source of atmospheric CO₂ (ref. 43) and there are currently intense efforts to establish the links between increasing global temperature and CO₂ concentrations with soil respiration and terrestrial carbon storage^{43,68}. Although there are indications that open-ocean *R* is at least as large as soil respiration, the uncertainty in the magnitude of the former is also much larger, and its importance as a source of CO₂ to the atmosphere is as yet unclear.

Current estimates of the magnitude of the oceanic sink of atmospheric CO₂ (approximately 2.2 Pg C yr⁻¹) based on net CO₂ uptake derived from global *p*CO₂ surveys⁶⁹, should not be affected by a reassessment of ocean *R*. However, consideration of oceanic respiration may help explain the distribution of CO₂ source regions in the ocean. In particular, the role of much of the subtropical ocean as a source of CO₂ to the atmosphere⁶⁹ is consistent with evidence that respiration in the photic layer exceeds primary production there^{10,11,14,15,53}. Yet, the role of the subtropical ocean as a CO₂ source has been exclusively attributed to temperature effects^{69,70}. Biological processes have been traditionally considered to act exclusively as

CO₂ sinks in the ocean^{41,64,69,70}, implicitly assuming that the oceanic *P/R* ratio consistently exceeds 1, an assumption that is unsupported by direct measurements of *P* and *R* (refs 10, 14, 15, 21). An increased awareness of the importance of oceanic *R* will probably lead to a better understanding of the drivers of this exchange, a key to eventually being able to predict how climate change will affect air–sea CO₂ flux. In the short-term, only planktonic respiration within the photic layer can influence air–sea CO₂ exchange directly, but the large amount of CO₂ released by planktonic respiration within the intermediate waters of the ocean certainly contributes to the CO₂ efflux to the atmosphere, as these waters ventilate within timescales of a few decades.

One of the main drivers of oceanic respiration is contemporary primary production, and climatic changes that affect patterns of ocean circulation and stratification, as well as increased atmospheric N inputs, may result in profound changes in global open-ocean productivity and in the balance between respiration and primary production^{3,71}. But factors other than recent production also strongly influence oceanic respiration, because respiration integrates organic matter inputs on a much larger time and spatial scale. This integration in turn provides long-term stability to the system. Temperature plays a major direct role in the regulation of respiration in aquatic ecosystems, so that global warming should lead to increased carbon consumption⁴⁴, as well as fundamental changes in the patterns of organic matter use by oceanic bacteria⁷². Although the net effect of these various processes on oceanic respiration and their consequences in terms of changes in carbon storage in the oceans are not well understood, a likely scenario is that global warming may lead to increased oceanic respiration, perhaps comparable to the predicted increase in soil respiration by 3.3 Gt C yr^{−1} per °C increase⁶⁸.

In summary, we identify oceanic respiration as one of the major components of the carbon flux in the biosphere, and highlight the considerable uncertainty in its magnitude as well as in our capacity to predict its response to global change. The International Panel for Climate Change (IPCC) has recently concluded that “current ocean climate models are severely limited by the ability to parameterize important biological activities and to specify the temporal and spatial variations in these parameterizations”. This is particularly true in the case of ocean respiration. Reducing the large present uncertainties as to the magnitude of open-ocean respiration will require the large-scale concerted effort of the international oceanographic community that led to our present knowledge of primary production in the oceans. After all, we cannot claim to grasp the global carbon cycle when we do not know whether the biota of the world’s oceans is a net source or sink for carbon. □

doi:10.1038/nature01165.

- Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
- Longhurst, A., Sathyendranath, S., Platt, T. & Caverhill, C. An estimate of global primary production in the ocean from satellite radiometer data. *J. Plankton Res.* **17**, 1245–1271 (1995).
- Ducklow, H. W. Ocean biogeochemical fluxes: New production and export of organic matter from the upper ocean. *Rev. Geophys.* **33** (Suppl.), 1271–1276 (1995).
- Balkanski, Y., Monfray, P., Batle, M. & Heimann, M. Ocean primary production derived from satellite data: An evaluation with atmospheric oxygen measurements. *Glob. Biogeochem. Cycles* **13**, 257–271 (1999).
- Morel, A. & Antoine, D. Small critters—big effects. *Science* **296**, 1980–1982 (2002).
- Williams, P. J. LeB. Microbial contribution to overall marine plankton metabolism: direct measurements of respiration. *Oceanolog. Acta* **4**, 359–364 (1981).
- Suess, E. Particulate organic carbon flux in the oceans—surface productivity and oxygen utilization. *Nature* **288**, 260–263 (1980).
- Karl, D. M., Hebel, D. V., Björkman, K. & Letelier, R. M. The role of dissolved organic matter release in the productivity of the oligotrophic North Pacific Ocean. *Limnol. Oceanogr.* **43**, 1270–1286 (1998).
- Lefèvre, D., Denis, M., Lambert, C. E. & Miguel, J.-C. Is DOC the main source of organic matter remineralization in the ocean water column? *J. Mar. Syst.* **7**, 281–291 (1996).
- Robinson, C. *et al.* Plankton respiration in the eastern Atlantic Ocean. *Deep-Sea Res.* **1** **49**, 787–813 (2002).
- Duarte, C. M., Aristegui, J., González, N., Agustí, S. & Anadón, R. Evidence for a heterotrophic subtropical NE Atlantic. *Limnol. Oceanogr.* **46**, 425–428 (2001).
- Harrison, W. G. *et al.* Basin-scale variability in plankton biomass and community metabolism in the subtropical North Atlantic Ocean. *Deep-Sea Res.* **II** **48**, 2241–2269 (2001).

- Gonzalez, N. *et al.* The metabolic balance of the planktonic community at the N. Atlantic Subtropical Gyre: The role of mesoscale instabilities. *Limnol. Oceanogr.* **46**, 946–952 (2001).
- del Giorgio, P. A., Cole, J. J. & Cimleris, A. Respiration rates in bacteria exceed plankton production in unproductive aquatic systems. *Nature* **385**, 148–151 (1997).
- Duarte, C. M. & Agustí, S. The CO₂ balance of unproductive aquatic ecosystems. *Science* **281**, 234–236 (1998).
- Williams, P. J. LeB. The balance of plankton respiration and photosynthesis in the open ocean. *Nature* **394**, 55–57 (1998).
- Duarte, C. M., Agustí, S., del Giorgio, P. A. & Cole, J. J. Regional carbon imbalances in the oceans. *Science* **284**, 1735b (1999).
- Biddanda, B. & Benner, R. Major contribution from mesopelagic plankton to heterotrophic metabolism in the upper ocean. *Deep-Sea Res.* **I** **44**, 2069–2085 (1997).
- Moriarty, D. J. W. & O’Donohue, M. J. Organic carbon transport from the Southern Ocean and bacterial growth in the Antarctic Intermediate Water masses of the Tasman Sea. *Mar. Ecol. Prog. Ser.* **119**, 291–297 (1995).
- Serret, P., Robinson, C., Fernández, E., Teira, E. & Tilstone, G. Latitudinal variation of the balance between plankton photosynthesis and respiration in the eastern Atlantic Ocean. *Limnol. Oceanogr.* **46**, 1642–1652 (2001).
- Aristegui, J. & Harrison, W. G. Decoupling of primary production and community respiration in the ocean: implications for regional carbon studies. *Aquat. Microb. Ecol.* (in the press).
- Pomeroy, L. R. & Johannes, R. E. Occurrence respiration of ultraplankton in the upper 500 meters of the ocean. *Deep-Sea Res.* **15**, 381–391 (1968).
- Menzel, D. W. & Ryther, J. H. Organic carbon and the oxygen minimum in the South Atlantic Ocean. *Deep-Sea Res.* **15**, 327–384 (1971).
- Vidal, M., Duarte, C. M. & Agustí, S. Dissolved organic nitrogen and phosphorus pools and fluxes in the central Atlantic Ocean. *Limnol. Oceanogr.* **44**, 106–115 (1999).
- Wyrtki, K. The oxygen minima in relation to ocean circulation. *Deep-Sea Res.* **9**, 11–23 (1962).
- Garfield, P. C., Packard, T. T., Friederich, G. E. & Codispoti, L. A. A subsurface particle maximum layer and enhanced microbial activity in the secondary nitrite maximum of the northeastern tropical Pacific Ocean. *J. Mar. Res.* **41**, 747–768 (1983).
- Packard, T. T. & Williams, P. J. LeB. Rates of respiratory oxygen consumption and electron transport in surface seawater from the northwest Atlantic. *Oceanolog. Acta* **4**, 351–358 (1981).
- Packard, T. T. Respiration and respiratory electron transport activity in plankton from the Northwest African upwelling. *J. Mar. Res.* **65**, 711–741 (1979).
- Boyd, P. W. *et al.* Transformations of biogenic particulates from the pelagic to the deep ocean realm. *Deep-Sea Res.* **II** **46**, 2761–2792 (1999).
- Doval, M. D. & Hansell, D. A. Organic carbon and apparent oxygen utilization in the western South Pacific and the central Indian Oceans. *Mar. Chem.* **68**, 249–264 (2000).
- Fiadeiro, M. E. & Craig, H. Three-dimensional modeling of tracers in the deep Pacific Ocean: I. Salinity and oxygen. *J. Mar. Res.* **36**, 323–355 (1978).
- Novoselov, A. A. Studies of oxygen consumption in the northern part of the Atlantic. *Okeanologiya* **2**, 84–92 (1962).
- Hansell, D. A. & Carlson, C. A. Deep-ocean gradients in the concentration of dissolved organic carbon. *Nature* **395**, 263–266 (1998).
- Jahnke, R. The global ocean flux of particulate organic carbon: Areal distribution and magnitude. *Glob. Biogeochem. Cycles* **10**, 71–88 (1996).
- King, F. D., Devol, A. H. & Packard, T. T. Plankton metabolic activity in the eastern tropical North Pacific. *Deep-Sea Res.* **25**, 689–704 (1978).
- Joiris, C. *et al.* A budget of carbon cycling in the Belgian coastal zone: relative roles of zooplankton, bacterioplankton and benthos in the utilization of primary production. *Netherlands J. Sea Res.* **16**, 260–275 (1982).
- Holligan, P. M., Williams, P. J. LeB., Purdie, D. & Harris, R. P. Photosynthesis, respiration and nitrogen supply of plankton populations in stratified, frontal and tidally mixed shelf waters. *Mar. Ecol. Prog. Ser.* **17**, 201–213 (1984).
- Christensen, J. P. & Packard, T. T. Oxygen utilization and plankton metabolism in a Washington Fjord. *Estuar. Coast. Mar. Sci.* **4**, 339–347 (1976).
- Hernández-León, S. *et al.* Large-scale and mesoscale distribution of plankton biomass and metabolic activity in the Northeastern Central Atlantic. *J. Oceanogr.* **55**, 471–482 (1999).
- del Giorgio, P. A. & Cole, J. J. *Marine Microbial Ecology* (ed. Kirchman, D.) 289–325 (Plenum, New York, 2000).
- Schlesinger, W. H. *Biogeochemistry: An Analysis of Global Change* (Academic, San Diego, 1991).
- Bauer, J. E., Williams, P. M. & Druffel, E. R. M. ¹⁴C activity of dissolved organic carbon fractions in the central North Pacific and Sargasso Sea. *Nature* **357**, 667–670 (1992).
- Woodwell, G. M. *et al.* Biotic feedbacks in the warming of the earth. *Clim. Change* **40**, 495–518 (1998).
- Pomeroy, L. R., Wiebe, W. J., Deibel, D., Thompson, R. J. & Rowe, G. T. Bacterial responses to temperature and substrate concentration during the Newfoundland spring bloom. *Mar. Ecol. Prog. Ser.* **75**, 143–159 (1991).
- Calbet, A. Mesozooplankton grazing effect on primary production: A global comparative analysis in marine ecosystems. *Limnol. Oceanogr.* **48**, 1824–1830 (2001).
- Pauly, D. & Christensen, V. Primary production required to sustain global fisheries. *Nature* **374**, 255–257 (1995).
- Jenkins, W. & Goldman, J. Seasonal oxygen cycling and primary production in the Sargasso Sea. *J. Mar. Sci.* **43**, 465–491 (1985).
- Duarte, C. M. & Cebrán, J. The fate of marine autotrophic production. *Limnol. Oceanogr.* **41**, 1758–1766 (1996).
- Bender, M., Ellis, T., Tans, P., Francey, R. & Lowe, D. Variability in the O₂/N₂ ratio of the southern hemisphere air, 1991–1994: Implications for the ocean carbon cycle. *Glob. Biogeochem. Cycles* **10**, 9–21 (1996).
- Carlson, C. A., Ducklow, H. W., Hansell, D. A. & Smith, W. O. Jr Organic carbon partitioning during spring phytoplankton blooms in the Ross Sea polynya and the Sargasso Sea. *Limnol. Oceanogr.* **43**, 375–386 (1998).
- Baines, S. B. & Pace, M. L. The production of dissolved organic carbon by phytoplankton and its importance to bacteria: Patterns across marine and freshwater systems. *Limnol. Oceanogr.* **36**, 1078–1090 (1991).

52. Cole, J. J., Findlay, S. & Pace, M. L. Bacterial production in fresh and saltwater ecosystems: a cross-system overview. *Mar. Ecol. Prog. Ser.* **43**, 1–10 (1988).
53. Hoppe, H.-G., Gocke, K., Koppe, R. & Begler, C. Bacterial growth and primary production along a north-south transect in the Atlantic Ocean. *Nature* **416**, 168–171 (2002).
54. Gattuso, J.-P., Franjignoulle, M. & Wollast, R. Carbon and carbonate metabolism in coastal aquatic ecosystems. *Annu. Rev. Ecol. Syst.* **29**, 405–433 (1998).
55. Mackenzie, F. T., Lerman, A. & Ver, L. M. B. Role of the continental margin in the global carbon balance during the past three centuries. *Geology* **26**, 423–426 (1998).
56. Bauer, J. E. & Druffel, E. R. M. Ocean margins as a significant source of organic matter to the deep open ocean. *Nature* **392**, 482–485 (1998).
57. Meybeck, M. Carbon, nitrogen and phosphorus transport by world rivers. *Am. J. Sci.* **282**, 401–450 (1982).
58. Duce, R. A. *et al.* The atmospheric input of trace species to the world ocean. *Glob. Biogeochem. Cycles* **5**, 193–259 (1991).
59. Cornell, S., Rendell, A. & Jickells, T. Atmospheric inputs of dissolved organic nitrogen to the oceans. *Nature* **376**, 243–246 (1995).
60. Kumar, N. *et al.* Increased biological productivity and export production in the glacial Southern Ocean. *Nature* **378**, 675–680 (1995).
61. Mopper, K. *et al.* Photochemical degradation of dissolved organic carbon and its impact on the ocean carbon cycle. *Nature* **353**, 60–62 (1991).
62. Cherrier, J., Bauer, J. E., Druffel, E. R. M., Coffin, R. B. & Chanton, J. P. Radiocarbon in marine bacteria: evidence for the ages of assimilated carbon. *Limnol. Oceanogr.* **44**, 730–736 (1999).
63. Christensen, J. P. Carbon export from continental shelves, denitrification and atmospheric carbon dioxide. *Continental Shelf Res.* **14**, 547–576 (1994).
64. Sundquist, E. T. The global carbon dioxide budget. *Science* **259**, 934–941 (1993).
65. Emerson, S. *et al.* Experimental determination of the organic carbon flux from open-ocean surface waters. *Nature* **389**, 951–954 (1997).
66. Sambrotto, R. N. *et al.* Elevated consumption of carbon relative to nitrogen in the surface ocean. *Nature* **363**, 248–250 (1993).
67. Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on ocean primary production. *Science* **281**, 200–206 (1998).
68. Raich, J. W., Potter, C. S. & Bhagawati, D. Interannual variability in global soil respiration, 1980–94. *Glob. Change Biol.* **8**, 800–812 (2002).
69. Takahashi, T. *et al.* Global sea-air CO₂ flux based on climatological surface ocean pCO₂, and seasonal biological and temperature effects. *Deep-Sea Res. II* **49**, 1601–1622 (2002).
70. Lefevre, N. & Taylor, A. Estimating pCO₂ from sea surface temperatures in the Atlantic gyres. *Deep-Sea Res. I* **49**, 539–554 (2002).
71. Smith, E. M. & Kemp, W. M. Seasonal and regional variations in plankton community production and respiration for the Chesapeake Bay. *Mar. Ecol. Prog. Ser.* **116**, 217–231 (1995).
72. Rivkin, R. B. & Legendre, L. Biogenic carbon cycling in the upper ocean: effects of microbial respiration. *Science* **291**, 2398–2400 (2001).

Acknowledgements We thank M. Pace, T. Bouvier and E. Smith for comments on the manuscript, and particularly P. leB. Williams for extensive input; we also thank H. Canut for encouragement and S. Agustí for inspiration. This work was supported by the Spanish Plan Nacional de Investigación y Desarrollo, the Cátedra Programme of the Banco de Bilbao y Vizcaya Foundation, and the US National Science Foundation.

Correspondence and requests for materials should be addressed to P.A.d.G. (e-mail: del_giorgio.paul@uqam.ca).

Recycled dehydrated lithosphere observed in plume-influenced mid-ocean-ridge basalt

Jacqueline Eaby Dixon*, Loretta Leist*, Charles Langmuir† & Jean-Guy Schilling‡

* RSMAS/MGG, University of Miami, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA

† Department of Earth & Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

‡ Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island 02882, USA

A substantial uncertainty in the Earth's global geochemical water cycle is the amount of water that enters the deep mantle through the subduction and recycling of hydrated oceanic lithosphere. Here we address the question of recycling of water into the deep mantle by characterizing the volatile contents of different mantle components as sampled by ocean island basalts and mid-ocean-ridge basalts. Although all mantle plume (ocean island) basalts seem to contain more water than mid-ocean-ridge basalts, we demonstrate that basalts associated with mantle plume components containing subducted lithosphere—'enriched-mantle' or 'EM-type' basalts—contain less water than those associated with a common mantle source. We interpret this depletion as indicating that water is extracted from the lithosphere during the subduction process, with greater than 92 per cent efficiency.

Quantifying the Earth's global geochemical water cycle is important for modelling physical properties of mantle materials¹, melt generation² and planetary evolution³. Water is added to oceanic crust at or near mid-ocean-ridge crests via hydrothermal alteration, and some or most of this water is lost during dehydration of subducting lithosphere, contributing to flux-melting of the overlying mantle wedge, generation of island arc magmas, and addition of water to the atmosphere–ocean system⁴. However, the amount of water that survives subduction and is transported to the deep mantle, perhaps contributing to subsequent 'wet' plume generation and ocean island volcanism⁵, is poorly known.

Recycling of crustal materials and mantle heterogeneity

Chemical diversity within the mantle is closely related to recycling of oceanic crust and sediments into the mantle by the plate tectonic process of subduction⁶. On the basis of radiogenic Pb, Nd and Sr isotopes in ocean island basalts, five isotopically extreme mantle components have been defined^{7–9}: (1) DMM (depleted MORB mantle); (2) HIMU ('high μ ' where $\mu = {}^{238}\text{U}/{}^{204}\text{Pb}$); (3) EM1 (enriched mantle 1); (4) EM2 (enriched mantle 2); and (5) LOMU ('low μ ') with very low time-integrated U/Pb ratio; MORB indicates mid-ocean-ridge basalt. In this context, 'enriched' refers to time-integrated Rb/Sr, Sm/Nd and/or (U + Th)/Pb ratios higher than primitive mantle (bulk silicate Earth). The origin and location of these end-members are actively debated, but it is generally accepted that the shallow upper mantle is the source of depleted MORB. HIMU is believed to represent recycled, old, hydrothermally altered oceanic crust, dehydrated and metamorphosed to eclogite during subduction. HIMU sources are depleted in Pb and K and enriched in U, Nb and Ta (refs 10, 11) relative to MORB. EM1 includes either recycled oceanic crust plus a few per cent pelagic sediment^{10–12} or metasomatized subcontinental lithosphere¹³. EM2 includes recycled oceanic crust containing a few per cent of continent-derived sediment¹¹. LOMU is distinguished from EM1 by its unusually high ${}^{87}\text{Sr}/{}^{86}\text{Sr}$ and ${}^{207}\text{Pb}/{}^{204}\text{Pb}$ ratios, and very low ${}^{143}\text{Nd}/{}^{144}\text{Nd}$ and ${}^{177}\text{Hf}/{}^{176}\text{Hf}$ ratios, suggesting an ancient continental source^{9,14,15}. Here we will refer to EM1, EM2, LOMU and mixtures of these simply as EM components, as all seem to include recycled materials of continental derivation.

Whereas the five end-members constrain the compositional

extremes, most data from ocean islands form linear mixing arrays that converge in a limited region of isotopic space that has been given various acronyms, including C ('common' mantle component; ref. 16) and FOZO ('focus zone'; refs 17, 18), used here. FOZO has moderately depleted Sr and Nd signatures, radiogenic Pb isotopes, and elevated ${}^3\text{He}/{}^4\text{He}$ ratios¹⁸. The high He isotope ratios of FOZO suggest an origin in the lower mantle^{16,19}. However, radiogenic isotopic compositions and trace-element ratios are inconsistent with FOZO being primitive mantle, and suggest involvement of a HIMU component^{6,10,20–25}.

It has also been proposed that EM components are associated with mantle metasomatism²⁶, implying concentrations of water and/or carbon dioxide greater than those in 'normal' mantle, if the metasomatic agent is a hydrous or carbonate phase. However, Dixon and Clague²⁷ showed that $\text{H}_2\text{O}/\text{Ce}$ in Loihi summit glasses correlates negatively with ${}^{206}\text{Pb}/{}^{204}\text{Pb}$, and interpreted this as evidence for a relatively 'dry' EM component in the Hawaiian plume. Here we report H_2O and CO_2 concentration data (measured by transmission infrared spectroscopy on MORB glasses from ridges influenced by various mantle plume components in the North and South Atlantic oceans); our aim is to see if the Hawaii result is unique, or if it represents a global EM characteristic.

$\text{H}_2\text{O}/\text{Ce}$ as a tracer of magmatic and nonmagmatic fractionations

Previous workers^{28,29} have found it useful to compare water concentrations to concentrations of elements or oxides of similar incompatibility, such as K_2O , P_2O_5 , La and Ce. Water has a bulk distribution coefficient ($D_{\text{H}_2\text{O}}$) of ~ 0.01 (refs 28, 29), slightly higher than K_2O and La and lower than Ce and P_2O_5 . For virtually all chemical species in oceanic basalts, ratios of more-incompatible to less-incompatible elements or oxides correlate positively. Thus H_2O ratioed to a slightly more compatible species (Ce, P_2O_5) correlates positively with other trace-element ratios indicative of source region enrichment or depletion (such as $(\text{La}/\text{Sm})_n$, $\text{K}_2\text{O}/\text{TiO}_2$), as has been observed at the Easter microplate (EMP) and at the Easter–Salas y Gomez seamount chain (ESC)³⁰, and for the East Pacific Rise and near-rise seamounts²⁹. These correlations are thought to represent mantle source characteristics, because as long as the extent of melting (F) is greater than the bulk distribution coefficient (D), trace-element ratios of incompatible elements in

basaltic melts will be similar to those of the source region. These data show that for areas such as the Easter microplate, water behaves as other incompatible elements and appears to be controlled by magmatic processes.

Shallow level processes introduce scatter into these correlations through degassing (which decreases H_2O/Ce) or assimilation of seawater-derived components (which increases H_2O/Ce). These modifications of water concentrations can be identified by examining

water concentration as a function of partial melting, differentiation, eruption depth, vesiculation, and chlorine concentration^{27–32}. For example, to correct for water degassing, bulk water concentrations can be estimated by adding the amount of water in vesicles to that dissolved in the glass^{27,31}.

Once shallow level processes have been accounted for, deeper processes can be investigated. For example, addition of hydrous fluid phases by mantle metasomatism, or addition of a mantle component containing hydrous phases, should increase H_2O/Ce , whereas incorporation of severely dehydrated components will decrease H_2O/Ce . If variations in H_2O/Ce correlate with radiogenic isotopic compositions, then the hydration or dehydration event probably occurred at the same time as the fractionation between parent and daughter elements.

Water in plume-influenced MORB from the South Atlantic

South Atlantic samples were dredged along the Mid-Atlantic Ridge between 40°S and 54°S during the RV *Maurice Ewing* EW93-09 cruise (see Supplementary Table S1). This region contains two distinct geochemical, gravity and bathymetric anomalies. The northern 'Discovery' anomaly is centred at 47.5°S and is influenced by the Discovery plume^{9,33}. The southern 'Shona' anomaly is centred at 51.5°S (refs 9, 33, 34).

We compare the South Atlantic samples to EMP and ESC glasses because both areas involve plume–ridge interaction, and Pacific and South Atlantic MORB have similar regional H_2O/Ce ratios²³, and thus similar depleted end-member compositions. Radiogenic isotopic data for EMP–ESC basalts (Fig. 1a) have been interpreted as the result of binary mixing between depleted MORB and mildly-HIMU FOZO (Salas y Gomez plume) sources^{35,36}.

Isotopic compositions of Discovery glasses show a wide range of $^{87}Sr/^{86}Sr$ at a relatively constant $^{206}Pb/^{204}Pb$, consistent with pseudo-binary mixing of DMM, FOZO and an EM component, specifically LOMU⁹ (Fig. 1a). Contribution of this EM component is least for samples 26, 28, 34 and 40, and greatest for sample 7, 9, 33 and 25 (ref. 37)³⁷. Whether this component is recycled in the upper or lower mantle is not known. However, the nearly solar $^{21}Ne/^{22}Ne$ ratios and high $^3He/^4He$ in some of the Discovery MORB glasses (samples 2–5) suggest recycling at least to the 660-km discontinuity, where more primitive mantle could have been entrained by the Discovery plume³⁸. Shona glasses form two groups in Sr–Nd–Pb–Hf isotopic space. Shona group 1 samples (dredges 11, 14, 15, 17 and 19) show mixing between DMM and FOZO with mild HIMU influence, similar to the Easter trend. Shona group 2 samples (dredges 18, 20, 21, 22 and 23) show additional influence of an EM component (EM1–EM2) thought to represent recycled oceanic crust + pelagic sediments¹⁵.

Water correlates positively with K_2O for the South Atlantic, but with very different slopes defined by Shona and Discovery samples (Fig. 1b). Linear correlations between H_2O and K_2O for Shona group 1, Easter, and other South Atlantic (31–46°S) MORB samples are indistinguishable (Fig. 1b). One outlier (sample 19) has more water at a given K_2O (about 0.1 wt% extra) than samples forming the main trend, with H_2O/Ce similar to several ESC samples, possibly related to serpentine assimilation during ascent²³. For the remaining samples, water is more compatible than K and less compatible than Ce, consistent with the behaviour of these elements in magmatic systems. Thus, the characteristics of the plume and DMM end-members for the Shona group 1 samples are similar to those of the EMP–ESC³⁰, for which the FOZO component has $H_2O/K_2O = \sim 0.8$ and $H_2O/Ce = 210 \pm 30$ and the DMM component has $H_2O/K_2O = \sim 5$ and $H_2O/Ce = 150 \pm 10$ (Fig. 1c).

Shona group 2 and Discovery samples ($^{87}Sr/^{86}Sr > 0.703$) that contain an EM component have distinctly lower H_2O concentrations and lower H_2O/Ce than Shona group 1 and Easter samples. Shona group 2 data form trends sub-parallel to Shona group 1 data but at progressively lower H_2O content and $^{206}Pb/^{204}Pb$. The

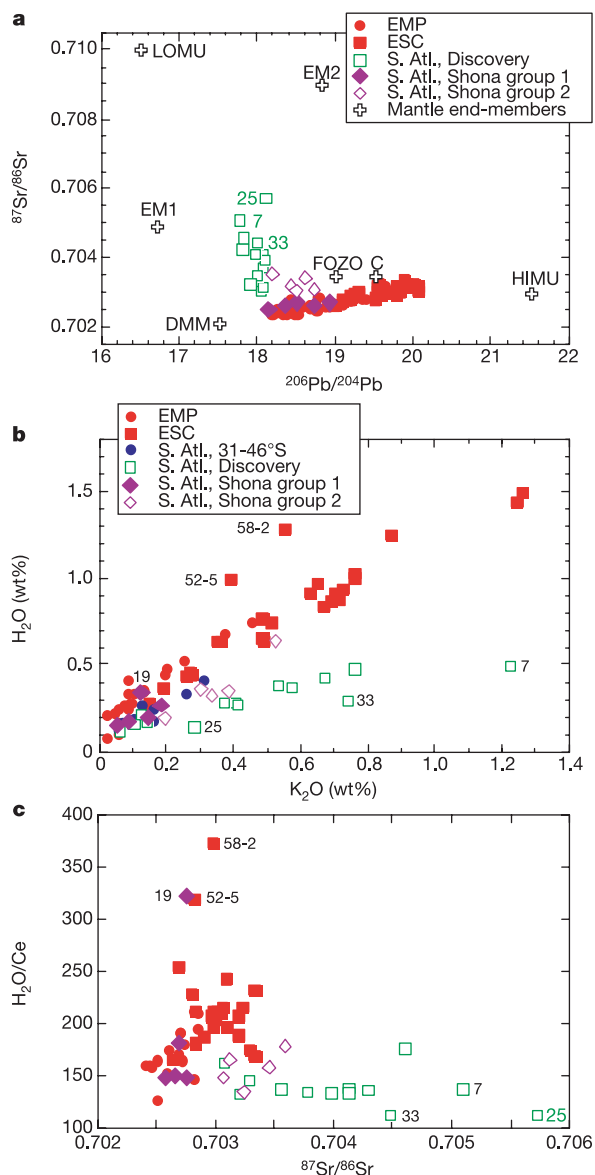


Figure 1 Comparison of radiogenic isotopic compositions, H_2O , K_2O and H_2O/Ce in basaltic glasses. Data are from the Discovery and Shona anomalies⁹ along the southern Mid-Atlantic Ridge, and from the Easter microplate and Easter–Salas y Gomez seamount chain (EMP–ESC) in the southeast Pacific^{35,36}. **a**, $^{87}Sr/^{86}Sr$ versus $^{206}Pb/^{204}Pb$. Isotopic compositions for mantle end-members are from ref. 51. Open symbols (Discovery and Shona group 2) represent samples having an EM-type mantle plume component. Filled symbols (Shona group 1 and EMP–ESC) represent samples containing a FOZO-type or mildly-HIMU plume component. **b**, H_2O versus K_2O . Water concentrations increase with increasing K_2O concentrations; however, EM-type basalts contain less water at a given K_2O than C-type basalts. **c**, H_2O/Ce versus $^{87}Sr/^{86}Sr$ for the South Atlantic and Easter microplate. EM-type basalts have significantly lower H_2O/Ce than FOZO-type basalts. Several samples have high H_2O/Ce ratios (>300) that do not correlate with major- or trace-element or isotopic compositions. These samples may have assimilated water during transport and eruption. Symbols are as in **b**.

Discovery samples form a distinct binary mixing line converging with that of the Easter group at the depleted source end (Fig. 1b, c). Discovery samples with the most definitive EM signature have $\text{H}_2\text{O}/\text{K}_2\text{O}$ half that of the Easter samples. The observation that sample 25 has relatively low water concentration and $\text{H}_2\text{O}/\text{Ce}$ is very important. Even though it has the highest $^{87}\text{Sr}/^{86}\text{Sr}$ ratio, reflecting a mantle source with time-integrated enrichment of incompatible elements (for example, high Rb/Sr), it is not the most enriched in incompatible elements, perhaps related to a recent episode of

melting and melt extraction³⁹. Sample 25 has only 0.15 wt% water. In order for this sample to have degassed significant quantities of water, it would have to have been erupted at or below the water solubility pressure of ~ 3 bar (30 m water depth), much shallower than its actual eruption depth at 2,032 m. It is almost certain that its low water concentration is a primary magmatic characteristic. Samples most enriched in $^{87}\text{Sr}/^{86}\text{Sr}$ (7, 25 and 33) have the lowest $\text{H}_2\text{O}/\text{K}_2\text{O}$. If these ratios were caused by the same type of (magmatic) process as observed for the Easter samples, then these samples should also have the highest $\text{H}_2\text{O}/\text{Ce}$. Instead, samples with the lowest $\text{H}_2\text{O}/\text{K}_2\text{O}$ have low $\text{H}_2\text{O}/\text{Ce}$, suggesting control by a different process that selectively lowers water content.

Thus, the EM components in the mantle sources for the Shona and Discovery samples are associated with lower water concentrations and lower ratios of water to similarly incompatible elements compared to the common plume component (FOZO).

Water in plume-influenced MORB from the North Atlantic

The north Mid-Atlantic Ridge samples were dredged by the RV *Atlantis II* (FAZAR expedition⁴⁰) between 33.2° and 40.5°N with depths varying between 926 and 3,900 m (see Supplementary Table S2). Two distinct geochemical anomalies occur within this region, including the long-wavelength anomaly associated with the intersection of the Azores archipelago with the Mid-Atlantic Ridge at $\sim 38\text{--}39^\circ\text{N}$ (Azores platform) and a short-wavelength anomaly centred at 35°N (the Great Meteor anomaly⁴¹, used here, or the 'north of Oceanographer fracture zone anomaly'⁴²). Sr and Pb isotopic data⁴³ (Fig. 2a) show that Azores platform basalts form a linear mixing array between DMM and FOZO. In contrast, basalts from the Great Meteor anomaly form a mixing array between FOZO and an EM component, with samples 9-10, 45-3 and 44-1 having the highest $^{87}\text{Sr}/^{86}\text{Sr}$ ratios. An exception is 49-3, which falls within the Azores platform array.

Because of the high vesicularity of many of the Azores platform and Great Meteor samples, loss of water into vesicles may have been significant. Both measured dissolved and calculated bulk water concentrations (see Supplementary Information) are shown in Fig. 2b and c. Samples from dredges 17, 19, 21 and 22 from the summit of the Azores platform (926–1,950 m) are highly vesicular (25–60 vol.%) and have lost about 10–30% of their initial water into vesicles. Dissolved water concentrations in the less vesicular Azores platform samples correlate positively with K_2O and have $\text{H}_2\text{O}/\text{Ce}$ ratios (253 ± 33) consistent with previously reported values for the North Atlantic²⁸. Estimated bulk $\text{H}_2\text{O}/\text{Ce}$ for the highly vesicular samples (247 ± 54) are similar to values for nonvesicular samples for all but the shallowest dredge (17), which may have suffered some gas loss (open-system degassing).

Great Meteor anomaly samples are also highly vesicular (5–60 vol.%), with vesicularity correlating roughly inversely with depth (5 vol.% at 3,900 m to 60 vol.% at 1,657 m). Samples with the largest EM component have dissolved and bulk water concentrations at a given K_2O (Fig. 2b) that are significantly lower than undegassed samples from the Azores platform. $\text{H}_2\text{O}/\text{Ce}$ is also 30% lower than the mean for the Azores platform samples (168 ± 29 versus 253 ± 33). Though the correlation is rough, extrapolation from FOZO through the Great Meteor data yields $\text{H}_2\text{O}/\text{Ce} < 100$ at $^{87}\text{Sr}/^{86}\text{Sr} > 0.705$. Thus, like the Discovery samples, EM components in the North Atlantic are depleted in water relative to FOZO.

Discussion

In each setting described above, there are two distinct mantle plume components: one close to FOZO (common mantle plume component), and the other enriched in recycled material of continental derivation (EM). In each case, the ratio of water to other similarly incompatible elements is lower in the EM component than in FOZO. Our data show that basalts influenced by EM mantle

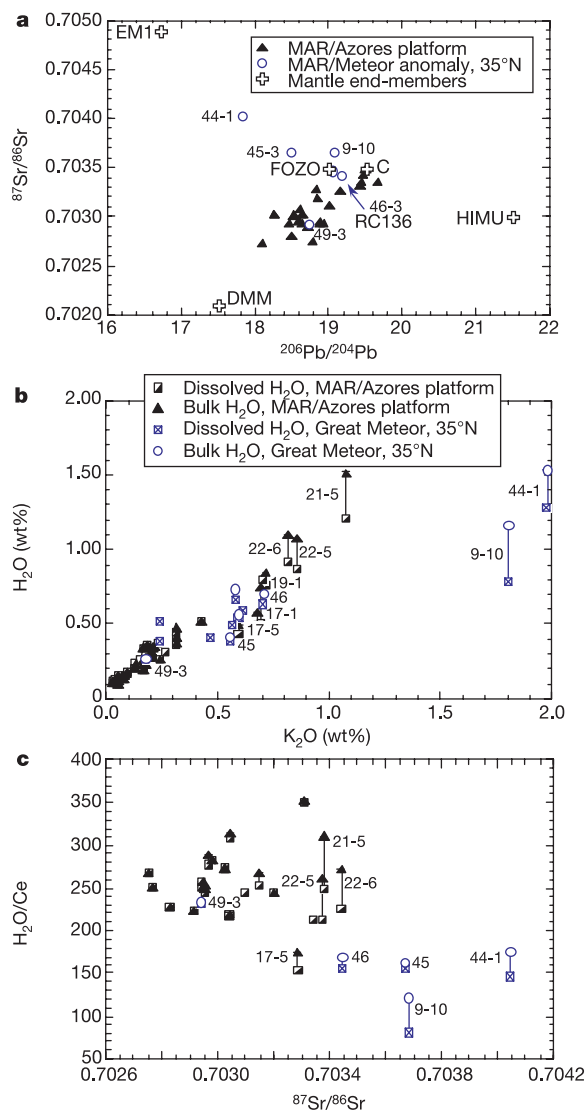


Figure 2 Comparison of radiogenic isotopic compositions, H_2O , K_2O and $\text{H}_2\text{O}/\text{Ce}$ in basaltic glasses from the North Atlantic. **a**, $^{87}\text{Sr}/^{86}\text{Sr}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$. Azores platform basalts⁴³ (black filled triangles) are consistent with mixing between DMM and C-type mantle components. Great Meteor anomaly basalts (blue open circles) are consistent mixing between EM-type and FOZO-type mantle components. **b**, H_2O versus K_2O . K_2O data from ref. 40. Measured dissolved H_2O concentrations are shown as black half-filled squares (Azores platform) and blue crossed squares (Great Meteor). Calculated bulk H_2O concentrations (see Supplementary Tables S1 and S2) are shown as black filled triangles (Azores platform) and blue open circles (Great Meteor). Lines connect dissolved and bulk water concentrations for individual samples. Calculated bulk H_2O concentrations for dredge 17 from the summit of the Azores platform are probably low owing to open-system degassing (J.E.D., manuscript in preparation). EM-type basalts have less water at a given K_2O content than FOZO-type basalts from the North Atlantic. **c**, $\text{H}_2\text{O}/\text{Ce}$ versus $^{87}\text{Sr}/^{86}\text{Sr}$ for the North Atlantic. EM-type basalts have lower $\text{H}_2\text{O}/\text{Ce}$ than FOZO-type basalts and depleted MORB. Symbols as in **b**.

components are depleted in water ('damp' plumes) relative to those influenced by FOZO ('wet' plumes).

What is FOZO?

We have shown that FOZO plumes are 'wetter' than EM plumes, but we still need to address the origin of H₂O in these mantle components. Though the origin of FOZO is an area of active debate, the following characteristics are generally accepted. First, FOZO is not pristine primitive mantle. Paradoxically, Sr- and Nd-isotopic compositions require moderate time-integrated depletion in incompatible elements (Rb/Sr lower and Sm/Nd higher than primitive mantle), whereas the present-day source is enriched in incompatible elements^{6,44}. In addition, Pb-, Os- and O-isotopic compositions and trace-element ratios (Nb/U, Nb/Th and Ce/Pb) suggest involvement of a HIMU component^{6,10,21–25,44}. Second, high concentrations of Ni, Cr and heavy rare-earth elements in FOZO require melting of a peridotitic source, and are inconsistent with melting of a purely eclogitic (HIMU) source^{44,45}. Third, FOZO typically has elevated ³He/⁴He, indicating involvement of mantle that is less degassed and more primitive than MORB mantle, probably located deeper than the MORB source^{17,18}. High ³He/⁴He also rules out generation of FOZO by mixing between DMM and HIMU components⁴⁵. Lastly, the globally widespread occurrence of FOZO in ocean island basalts and carbonatites makes it unlikely that FOZO is the product of melting heterogeneous recycled crustal materials that make up only 4–10% of mantle volume^{18,45}. Similar to a recent model by Donnelly *et al.*⁴⁴, we suggest that FOZO represents dominantly primitive or slightly depleted peridotitic mantle (~97–99%), metasomatized by a small amount (1–3%) of a small extent melt (2–6%) of a HIMU component. If the lower mantle is composed of peridotite with 'streaks' of recycled eclogitized oceanic crust (HIMU), then this metasomatism could occur during ascent of plumes, a process recently proposed to explain major-element heterogeneity in picrites and komatiites⁴⁶. In such a mixture, concentrations of noble gases and compatible elements (Ni, Cr and heavy rare-earth elements) in the FOZO source come almost completely from the peridotitic mantle, whereas roughly half to two-thirds of the highly incompatible elements are supplied by the metasomatic HIMU-melt. Thus, water in FOZO plumes is a combination of primitive and recycled water.

H₂O in mantle sources of FOZO and HIMU plumes

On the basis of eclogite measurements⁴⁷, subducted oceanic crust

(HIMU source) is low in water, with ~600 p.p.m. H₂O, ~6 p.p.m. Ce, and H₂O/Ce = ~100. These data suggest that subducted oceanic crust is efficiently (~97%) dehydrated during subduction (Table 1). In contrast, mantle sources of FOZO plumes have ~750 p.p.m. H₂O and H₂O/Ce of ~210 in the Pacific and South Atlantic and ~250 in the North Atlantic (ref. 25, and P. Asimow, C.L. and J.E.D., manuscript in preparation). The high H₂O/Ce in FOZO plumes cannot be derived from recycled oceanic crust; a significant amount of water must be juvenile, left over from planetary accretion. We believe that it is the presence of this juvenile water that makes FOZO plumes 'wet'.

Water in EM components

We have shown that recycled oceanic crust is efficiently dehydrated during the subduction process. The lower H₂O/Ce of EM plumes relative to FOZO plumes indicates that recycled continentally derived components are also efficiently dehydrated. For example, an estimate of globally subducted sediment⁴⁸ contains 57 p.p.m. Ce and ~7.3 wt% H₂O (H₂O/Ce = 1,280). Reduction of the H₂O/Ce to <100, lower than the ambient upper mantle of >150, requires >92% dehydration of the water from sediments (Table 1). These dehydrated sediments are still fairly hydrous (<0.57 wt% H₂O), but addition of only a few per cent sediments to HIMU will only increase the water concentration by 50 to 150 p.p.m. Thus our data suggest that dehydration is efficient for all recycled materials. These estimates of dehydration are consistent with other approaches that have considered mass balance at the subduction zone itself⁴⁹.

Discovery of dense hydrous magnesian silicate phases in laboratory experiments is generally considered consistent with a scheme whereby water may be efficiently transported into the deep mantle during subduction⁵⁰. However, laboratory generation of phases capable of holding significant amounts of water does not constitute evidence that these phases in fact carry significant water to the deep mantle. Our data suggest that much of the recycled material that has passed through the subduction 'factory' has H₂O/Ce < 100. Thus, as is the case for Pb (refs 21–23), the subduction process effectively partitions water into the exosphere (mantle wedge, crust, ocean and atmosphere), resulting in time-integrated depletion of water relative to other incompatible elements in the recycled lithosphere and accompanying continental-derived materials. Thus, like a well-squeezed sponge, the residual slab and EM plume components are not 'wet', but 'damp'. □

Received 28 August; accepted 9 October 2002; doi:10.1038/nature01215.

Table 1 Dehydration efficiency

Rock	Ce (p.p.m.)‡	H ₂ O/Ce	H ₂ O (wt%)	Dehydration (%)
Crustal components: Pre-subduction				
Atlantic MORB*	12	250	0.3	
Pacific MORB*	14	150	0.2	
Mature oceanic crust*	~6	2,500–5,000	2–3	
Global subducted sediment†	57	1,280	7.3	
Crustal components: Post-subduction				
Mature oceanic crust	~6	~100	0.06§	97
Global subducted sediment	57	<100	<0.57	>92
Mantle components				
DMM*	0.5	200	0.010	
Atlantic FOZO	3	250	0.075	
Pacific FOZO	3.8	200	0.075	
EM#	4	<100	<0.04	

* GERM (Geochemical Earth Reference Model) database (<http://earthref.org/databases/GERMRD/main.htm>).

† GLOSS (Global Subducting Sediment)⁴⁸.

‡ Represents conservative estimate—some rare-earth elements are lost from subducting lithosphere during subduction.

§ Bulk water concentration in eclogite⁴⁷.

|| Asimow, P., C.L. and J.E.D., manuscript in preparation.

¶ From refs 27 and 30.

EM sources have Ce contents similar to HIMU sources⁶ and slightly greater than FOZO-type sources. EM mantle sources have roughly half as much water as FOZO-type sources (this study).

- Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
- Hirose, K. & Kawamoto, T. Hydrous partial melting of lherzolite at 1 GPa: the effect of H₂O on the genesis of basaltic magmas. *Earth Planet. Sci. Lett.* **133**, 463–473 (1995).
- Abe, Y., Ohtani, E., Okuchi, T., Richter, K. & Drake, M. *Origin of the Earth and Moon* (eds Canup, R. M. & Richter, K.) 413–433 (Univ. Arizona Press, Tucson, 2000).
- Schmidt, M. W. & Poli, S. Experimental base water budgets for dehydrating slabs and consequences for arc magma generation. *Earth Planet. Sci. Lett.* **163**, 361–379 (1998).
- Schilling, J.-G., Bergeron, M. B. & Evans, R. Halogens in the mantle beneath the North Atlantic. *Phil. Trans. R. Soc. Lond. A* **297**, 147–178 (1980).
- Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
- White, W. M. Sources of oceanic basalts: Radiogenic isotopic evidence. *Geology* **13**, 115–118 (1985).
- Zindler, A. & Hart, S. Chemical geodynamics. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
- Douglass, J., Schilling, J.-G. & Fontignie, D. Plume-ridge interactions of the Discovery and Shona mantle plumes with the southern Mid-Atlantic Ridge (40°–55°S). *J. Geophys. Res.* **104**, 2941–2962 (1999).
- Weaver, B. L. The origin of ocean island basalt end-member compositions: Trace element and isotopic constraints. *Earth Planet. Sci. Lett.* **204**, 381–397 (1991).
- Chauvel, C., Hofmann, A. W. & Vidal, P. HIMU-EM: the French Polynesian connection. *Earth Planet. Sci. Lett.* **110**, 99–119 (1992).
- Blichert-Toft, J., Frey, F. A. & Albarède, F. Hf isotope evidence for pelagic sediments in the source of Hawaiian basalts. *Science* **285**, 879–882 (1999).
- McKenzie, D. & O'Nions, R. K. Mantle reservoirs and ocean island basalts. *Nature* **301**, 229–231 (1983).
- Kamenetsky, V. S. *et al.* Remnants of Gondwanan continental lithosphere in oceanic upper mantle: Evidence from the South Atlantic Ridge. *Geology* **29**, 243–246 (2001).
- Andres, M., Blichert-Toft, J. & Schilling, J.-G. Hafnium isotopes in basalts from the Southern Mid-Atlantic Ridge from 40°S to 55°S: Trace element evidence for contamination of ambient asthenospheric mantle. *Earth Planet. Sci. Lett.* (in the press).

16. Hanan, B. B. & Graham, D. W. Lead and helium isotope evidence from oceanic basalts for a common deep source of mantle plumes. *Science* **272**, 991–995 (1996).
17. Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. Mantle plumes and entrainment: The isotopic evidence. *Science* **256**, 517–520 (1992).
18. Hauri, E. H., Whitehead, J. A. & Hart, S. R. Fluid dynamic and geochemical aspects of entrainment in mantle plumes. *J. Geophys. Res.* **99**, 24275–24300 (1994).
19. Allegre, C. J., Hofmann, A. W. & O'Nions, R. K. The argon constraints on mantle structure. *Geophys. Res. Lett.* **23**, 3555–3557 (1996).
20. Hofmann, A. W. Nb in Hawaiian magma: constraints on source composition and evolution. *Chem. Geol.* **57**, 17–30 (1986).
21. Hofmann, A. W., Jochum, K. P., Seufert, M. & White, W. M. Nb and Pb in oceanic basalts: new constraints on mantle evolution. *Earth Planet. Sci. Lett.* **79**, 33–45 (1986).
22. Miller, D. M., Goldstein, S. L. & Langmuir, C. H. Cerium/lead and lead isotope ratios in arc magmas and the enrichment of lead in the continents. *Nature* **368**, 514–520 (1994).
23. Chauvel, C., Goldstein, S. L. & Hofmann, A. W. Hydration and dehydration of oceanic crust controls Pb evolution in the mantle. *Chem. Geol.* **126**, 65–75 (1995).
24. Shirey, S. B. & Walker, R. J. The Re–Os isotope system in cosmochemistry and high-temperature geochemistry. *Annu. Rev. Earth Planet. Sci.* **26**, 423–500 (1998).
25. Eiler, J. M. *Stable Isotopes in Igneous Processes* (eds Cole, D. and Valley, J. W.) (Mineralogical Society of America, Washington DC, 2002).
26. Menzies, M. & Murthy, V. R. Enriched mantle: Nd and Sr isotopes in diopsides from kimberlite nodules. *Nature* **283**, 634–636 (1980).
27. Dixon, J. E. & Clague, D. A. Volatiles in basaltic glasses from Loihi Seamount, Hawaii: Evidence for a relatively dry plume component. *J. Petrol.* **42**, 627–654 (2001).
28. Michael, P. J. Regionally distinctive sources of depleted MORB: Evidence from trace elements and H₂O. *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
29. Danyushevsky, L. V., Eggins, S. M., Falloon, T. J. & Christie, D. M. H₂O abundance in depleted to moderately enriched mid-ocean ridge magmas; Part I: Incompatible behaviour, implications for mantle storage, and origin of regional variations. *J. Petrol.* **41**, 1329–1364 (2000).
30. Simons, K., Dixon, J. E., Schilling, J.-G., Kingsley, R. & Poreda, R. Volatiles in basaltic glasses from the Easter-Salas y Gomez Seamount Chain and Easter Microplate: Implications for geochemical cycling of volatile elements. *Geochim. Geophys. Geosyst.* **3**, doi: 10.1029/2001GC000173 (2002).
31. Dixon, J. E., Clague, D. A., Wallace, P. & Poreda, R. Volatiles in alkalic basalts from the North Arch Volcanic Field, Hawaii: Extensive degassing of deep submarine-erupted alkalic series lavas. *J. Petrol.* **38**, 911–939 (1997).
32. Michael, P. J. & Cornell, W. C. Influence of spreading rate and magma supply on crystallization and assimilation beneath mid-ocean ridges: Evidence from chlorine and major element chemistry of mid-ocean ridge basalts. *J. Geophys. Res.* **103**, 18325–18356 (1998).
33. Douglass, J., Schilling, J.-G. & Kingsley, R. H. Influence of the Discovery and Shona mantle plumes on the southern Mid-Atlantic Ridge: Rare earth evidence. *Geophys. Res. Lett.* **22**, 2893–2896 (1995).
34. le Roex, A. P., Dick, H. J. B., Gulen, L., Reid, A. M. & Erlank, A. J. Local and regional heterogeneity in MORB from the Mid-Atlantic Ridge between 54.5°S and 51°S: Evidence for geochemical enrichment. *Geochim. Cosmochim. Acta* **51**, 541–555 (1987).
35. Kingsley, R. H. & Schilling, J.-G. Plume-ridge interaction in the Easter-Salas y Gomez seamount chain-Easter Microplate system: Pb isotope evidence. *J. Geophys. Res.* **103**, 24159–24177 (1998).
36. Cheng, Q. C., Macdougall, J. D. & Zhu, P. Isotopic constraints on the Easter Seamount Chain source. *Contrib. Mineral. Petrol.* **135**, 225–233 (1999).
37. Douglass, J. & Schilling, J.-G. Systematics of three-component, pseudo-binary mixing lines in 2D isotope ratio space representations and implications for mantle plume-ridge interaction. *Chem. Geol.* **163**, 1–23 (2000).
38. Sarda, P., Moreira, M., Staudacher, T., Schilling, J.-G. & Allègre, C. J. Rare gas systematics on the southernmost Mid-Atlantic Ridge: Constraints on the lower mantle and the Dupal source. *J. Geophys. Res.* **105**, 5973–5996 (2000).
39. le Roux, P. J. *et al.* Mantle heterogeneity beneath the southern Mid-Atlantic Ridge: trace element evidence for contamination of ambient asthenospheric mantle. *Earth Planet. Sci. Lett.* (in the press).
40. Langmuir, C. H., Klinkhammer, G. & Bougalult, H. *Cruise Report: Evaluation of the Relationships among Segmentation, Hydrothermal Activity and Petrological Diversity on the Mid-Atlantic Ridge (FAZAR)* Technical Report LDEO-92-3 (Lamont-Doherty Earth Observatory, Columbia Univ., Palisades, 1992).
41. Rideout, M. L. & Schilling, J.-G. Rare-earth elements, ⁸⁷Sr/⁸⁶Sr, and ¹⁴³Nd/¹⁴⁴Nd mantle source variations. *Init. Rep. DSDP* **82**, 483–496 (1985).
42. Shirey, S. B., Bender, J. F. & Langmuir, C. H. Three-component isotopic heterogeneity near the Oceanographer transform, Mid-Atlantic Ridge. *Nature* **325**, 217–223 (1987).
43. Dosso, L. *et al.* The age and distribution of mantle heterogeneity along the Mid-Atlantic Ridge (31–41°N). *Earth Planet. Sci. Lett.* **170**, 269–286 (1999).
44. Donnelly, K., Langmuir, C. H., Goldstein, S. L. & Lagatta, A. The origin of alkali and ocean island basalts: contradictions and solutions. *Eos* **82**, F1402 (2001).
45. Bell, K. & Tilton, G. R. Probing the mantle: The story from carbonatites. *Eos* **83**, 273–277 (2002).
46. Gibson, S. A. Major element heterogeneity in Archean to Recent mantle plume starting-heads. *Earth Planet. Sci. Lett.* **195**, 59–74 (2002).
47. Nadeau, S., Philippot, P. & Pineau, F. Fluid inclusion and mineral isotopic compositions (H–C–O) in eclogitic rocks as tracers of local fluid migration during high-pressure metamorphism. *Earth Planet. Sci. Lett.* **114**, 431–448 (1993).
48. Plank, T. & Langmuir, C. H. The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chem. Geol.* **145**, 325–394 (1998).
49. Peacock, S. M. Fluid processes in subduction zones. *Science* **248**, 329–337 (1990).
50. Thompson, A. B. Water in the Earth's upper mantle. *Nature* **358**, 295–302 (1992).
51. Schilling, J.-G. *Les Isotopes Radiogéniques en Géologie: Chronologie Géochimie* (ed. Duthou, J. L.) 1–34 (Société Française de Minéralogie et de Cristallographie, Paris, 1992).

Supplementary Information accompanies the paper on *Nature's* website
 (► <http://www.nature.com/nature>).

Acknowledgements This research was funded by the NSE. A review by N. Arndt improved the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.D.
 (e-mail: jdixon@rsmas.miami.edu).

Downward pumping of magnetic flux as the cause of filamentary structures in sunspot penumbrae

John H. Thomas^{*†}, Nigel O. Weiss^{*}, Steven M. Tobias[‡]
& Nicholas H. Brummell[§]

^{*} Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK

[†] Department of Physics and Astronomy and Department of Mechanical Engineering, University of Rochester, Rochester, New York 14627-0171, USA

[‡] Department of Applied Mathematics, The University of Leeds, Leeds LS2 9JT, UK

[§] JILA and Department of Astrophysical and Planetary Sciences, University of Colorado, Boulder, Colorado 80309-0440, USA

The structure of a sunspot is determined by the local interaction between magnetic fields and convection near the Sun's surface^{1,2}. The dark central umbra is surrounded by a filamentary penumbra, whose complicated fine structure has only recently been revealed by high-resolution observations^{3–14}. The penumbral magnetic field has an intricate and unexpected interlocking-comb structure and some field lines, with associated outflows of gas¹⁵, dive back down below the solar surface at the outer edge of the spot. These field lines might be expected to float quickly back to the surface because of magnetic buoyancy, but they remain submerged. Here we show that the field lines are kept submerged outside the spot by turbulent, compressible

convection, which is dominated by strong, coherent, descending plumes^{16,17}. Moreover, this downward pumping of magnetic flux explains the origin of the interlocking-comb structure of the penumbral magnetic field, and the behaviour of other magnetic features near the sunspot.

Near the solar surface there is a shallow, strongly superadiabatic layer with vigorous small-scale convection, lying over a weakly unstable region of gentler, larger-scale convection. The small-scale motion is responsible for the observed pattern of short-lived convection cells with diameters around 1,000 km (granules) at the solar surface, while the larger-scale motion manifests itself as supergranules with diameters around 30,000 km. We propose the picture of a sunspot shown in Fig. 1, in which granular convection plays a key role in submerging the returning penumbral flux tubes and establishing the structure of the penumbral magnetic field.

The basic ideas of convective pumping of magnetic flux may be traced back to earlier concepts (flux expulsion, turbulent diamagnetism, and topological pumping—see references in ref. 18). Recent numerical experiments have demonstrated that turbulent pumping of magnetic flux by penetrative convection at the base of the convection zone is an important ingredient of the dynamo that is responsible for cyclic activity in the Sun^{18–22}. These simulations of three-dimensional, compressible, turbulent convection have revealed the true dynamical nature of the pumping mechanism. The vigorous sinking plumes transport magnetic flux preferentially downwards out of the turbulent convecting region and into a stably stratified region below, where the flux can be amplified and stored. The calculations show that the pumping mechanism is remarkably robust, and works well even when the convective stability of the

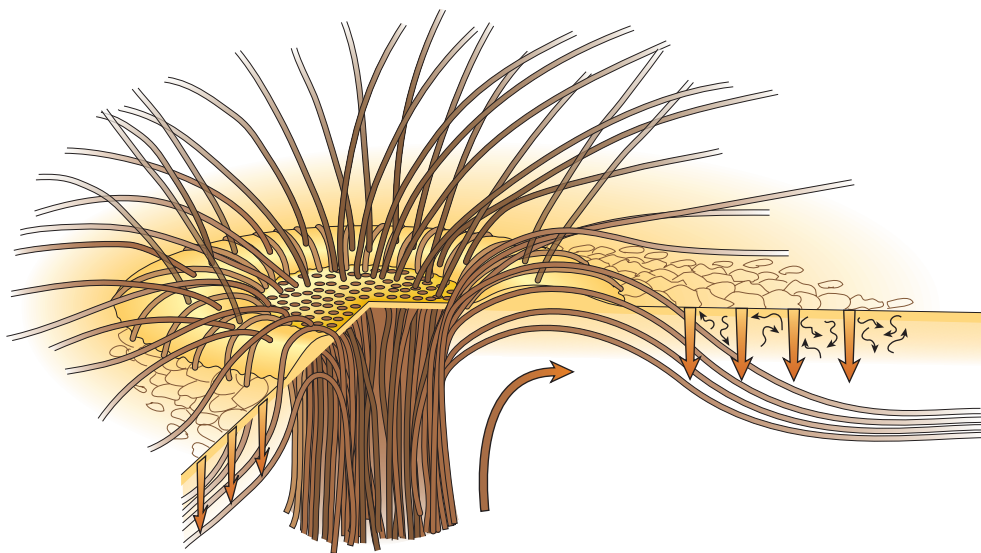


Figure 1 Sketch showing the interlocking-comb structure of the magnetic field in the filamentary penumbra of a sunspot. The bright radial filaments, where the magnetic field is inclined (at about 40° to the horizontal in the outer penumbra), alternate with dark filaments in which the field is nearly horizontal^{3,5–7,13}. Within the dark filaments, some magnetic flux tubes (that is, bundles of magnetic field lines) extend radially outward beyond the penumbra along an elevated magnetic canopy while other, 'returning' flux tubes dive back down below the surface. The sunspot is surrounded by a layer of small-scale granular convection (thin squiggly black arrows) embedded in the radial outflow (thick curved brown arrow) associated with a long-lived annular supergranule (the moat cell). The submerged parts of the returning flux tubes are held down by turbulent pumping (indicated by thick vertical brown arrows) due to granular convection in the moat. There is also a persistent horizontal outflow in the penumbra (the Evershed flow), which is mostly

confined to thin, nearly horizontal, radial channels within the dark filaments. Because of the relatively high electrical conductivity of the gas, this flow is constrained to be along magnetic field lines. A small fraction of the flow runs along field lines that extend radially outward beyond the penumbral boundary along the magnetic canopy, elevated slightly above the surrounding quiet photosphere. Most of the Evershed flow, however, runs along arched magnetic flux tubes that dive back down below the visible surface at points either just within or just outside the outer boundary of the penumbra^{8–12}. This configuration is in close agreement with the siphon-flow model, in which a flow along an arched magnetic flux tube is driven by a drop in gas pressure between the two footpoints of the arch¹⁵; at the outer footpoint the intergranular magnetic field is stronger, the magnetic pressure is higher and the gas pressure is therefore lower.

lower layer is reduced¹⁸. We therefore expect that downward pumping by the granular convection just beneath the solar surface will be particularly effective.

To test the effects of flux pumping on sunspot structure, we have carried out calculations that are specifically designed to model turbulent pumping by granular convection. The model consists of a shallow, strongly superadiabatic layer representing the region where granules form, lying over a deep, nearly adiabatic layer representing the region of weaker, larger-scale (such as supergranular) motion. For computational reasons, the lower layer is in fact weakly stable, but this does not affect the qualitative features of the results. Figure 2 shows the initial and final states for this flux-pumping calculation. First, a purely hydrodynamic simulation with no magnetic field¹⁷ is carried out until a statistically steady state is reached. This steady state takes the form of vigorous, turbulent convection above a largely quiescent layer (Fig. 2a). The convection is dominated by a network of downflows, with the strongest flows in the downward-sinking plumes that penetrate into the layer below. Then, at time $t = 0$, we introduce into the fully developed convection a thin uniform layer of horizontal magnetic field $\mathbf{B} = (0, B_y, 0)$, as shown in Fig. 2b, and adjust the density of the layer so that the pressure remains continuous, although the magnetic slab is nearly evacuated and hence very buoyant.

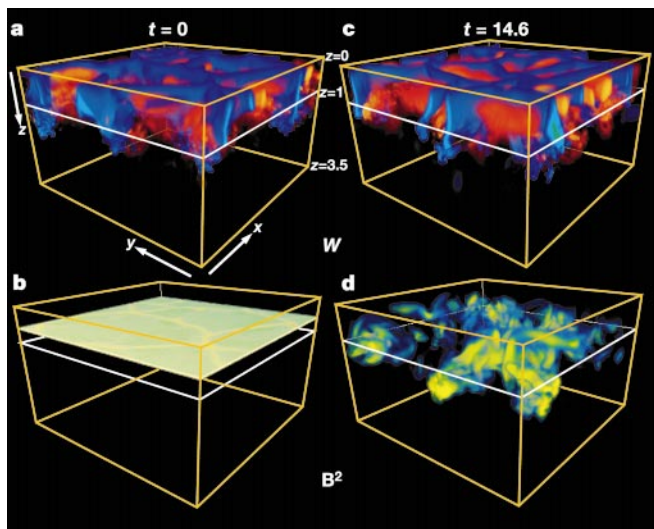


Figure 2 Downward pumping of magnetic flux by turbulent granular convection. Shown are volume renderings of the results of our numerical simulation, in a rectangular box of horizontal dimensions 6×6 and vertical dimension 3.5 in units of the depth of the upper 'granulation' layer. The vertical scale is exaggerated. The boundary conditions are periodic in the horizontal directions and stress-free at the top and bottom of the box; the magnetic field $\mathbf{B}$ is constrained to be vertical at the upper boundary ($B_x = B_y = 0$ at $z = 0$) but no magnetic flux is allowed to escape across the bottom boundary ($\partial B_x / \partial z = \partial B_y / \partial z = 0$ at $z = 3.5$). The Rayleigh number $Ra = 5 \times 10^5$ and the initial field strength corresponds locally to a Chandrasekhar number $Q = 10^5$. The stiffness parameter $S = (m_2 - m_{ad}) / (m_{ad} - m_1)$, where m_1 and m_2 are the polytropic indices of the upper (unstable) and lower (stable) layer, respectively, and the adiabatic index $m_{ad} = 3/2$. Here we set $m_1 = 1$, $m_2 = 1.75$ so that $S = 0.5$. Other parameters are as in ref. 18, where the corresponding calculation had $S = 15$. **a**, Vertical velocity w of fully developed nonmagnetic convection at time $t = 0$, rendered with red/yellow indicating weak/strong upflows and blue/light blue indicating weak/strong downflows. **b**, The magnetic energy density just after the slab of uniform magnetic field (in the y direction) is introduced at $t = 0$. **c**, The vertical velocity after the pumping phase (at a dimensionless time $t = 14.6$). **d**, The magnetic energy density (increasing as blue–green–yellow) after the pumping phase. Strong coherent downflows penetrate from the 'granular' layer into the 'supergranular' layer below. These strong, sinking plumes amplify the magnetic energy locally, while pumping magnetic flux downward into the stable layer.

The magnetic field is immediately advected by the convection. Although some of the magnetic flux is transported upwards by the weaker upflows and magnetic buoyancy, and the upper boundary condition allows flux to leak through the surface, the strong downflows eventually overcome these effects and carry flux down towards the stable layer, amplifying the local magnetic energy as they plunge down. The velocity at the end of this pumping phase is shown in Fig. 2c, with the corresponding state of the magnetic energy shown in Fig. 2d. Although some magnetic flux remains within the convecting layer, a larger proportion has been swept down by the penetrating plumes into the layer below.

Three aspects of the final state are displayed in Fig. 3. The enstrophy (vorticity squared) is strongly concentrated in the vigorous sinking plumes at the corners of the network of downflows. These thin plumes splash down and broaden as they penetrate into the nearly adiabatic layer below. The field lines show the different field topologies in the stable and unstable regions. In the upper region, the plumes advect the tangled magnetic field, twisting and carrying it down into the nearly adiabatic layer to achieve the net transport downward of magnetic flux. In the lower region, the field is predominantly in the y direction, though the interaction with convection is still significant.

This downward transport is shown quantitatively in Fig. 4. In the initial pumping phase, the flux is transported predominantly by the action of convection. After an initial rise and loss of flux owing to magnetic buoyancy, efficient pumping of the magnetic flux takes over and the maximum of the flux appears in the nearly adiabatic layer (below $z = 1$). Although a significant mean flux remains in the turbulent layer, most of the net flux is in the stable region. This is the key feature of the computation—the downward transport of magnetic flux out of the layer of vigorous convection into the layer below. In the second phase of the calculation, flux drifts downward by diffusion (which would be a small factor in the Sun, but is significant in our model) and the flux profile diffuses deeper into the interior.

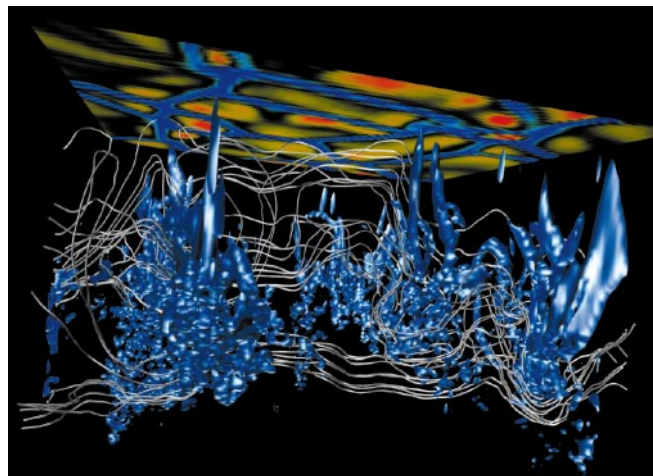


Figure 3 Flux pumping by vigorous sinking plumes. Perspective view from below the surface, showing an illuminated surface (in blue) of constant enstrophy density (that is, the square of the vorticity), which outlines the regions of strong descending plumes. At the top of the figure is a two-dimensional representation of the vertical velocity just below the surface, showing the small-scale convection as a network of downflows surrounding broad upflows, with yellow/red indicating weak/strong upflows, and light blue/blue indicating weak/strong downflows. The magnetic field is visualized by tracing selected bundles of representative field lines, some in the upper 'granulation' layer and some in the slightly stable, nearly adiabatic layer below. Note that these field lines correspond to different field strengths, and give no indication of the strength of the field. The tangled field in the upper region is pumped downwards into the stably stratified layer, where the field is predominantly in the y direction.

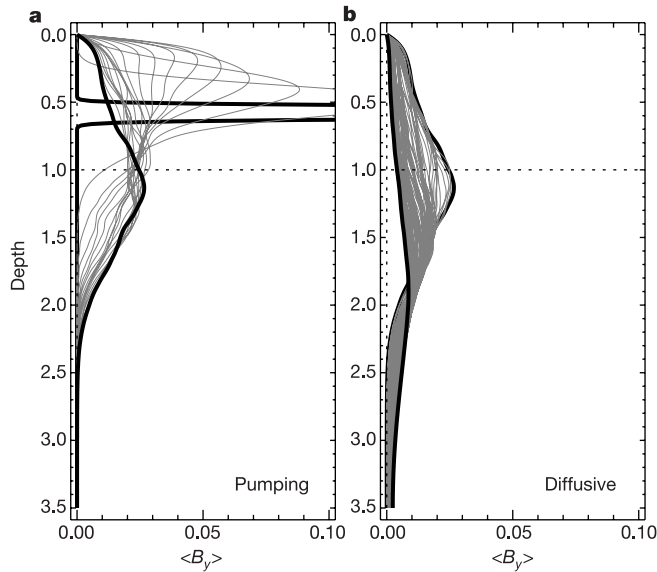


Figure 4 Downward pumping of magnetic flux in the numerical simulation. Plots show the horizontal average of the y -component of the magnetic field as a function of depth z at evenly spaced times, for two phases of the calculation. Panel **a** shows the rapid pumping phase ($0 \leq t \leq 9$), in which redistribution of the field occurs mainly as a result of the convective motions, and panel **b** shows the later slow diffusive phase, in which magnetic diffusion is the main transport mechanism. Heavy lines denote the flux distributions at the beginning and end of each phase. At first magnetic flux moves upwards, owing to magnetic buoyancy, and some leaks out through the upper boundary. Then the mean field is rapidly pumped downwards into the weakly stable region. Note, however, that a strong tangled field still survives in the vigorously convecting layer. The normalized field strength B can be related to the local gas pressure in the original static atmosphere by introducing the plasma β (the ratio of gas pressure to magnetic pressure): at $z = 0.5$, $\beta \approx 14/B^2$, while at $z = 1.0$, $\beta \approx 48/B^2$.

Overall, this simulation implies that the vigorously convecting granulation layer at the solar surface is very effective at expelling magnetic flux and pumping some of this flux into the underlying layer of gentler supergranular and larger-scale flows, where magnetic buoyancy becomes relatively more powerful. This in turn implies that flux pumping is an important mechanism for the formation and maintenance of a sunspot penumbra. A sunspot forms through the coalescence of small pores without penumbrae. As the total magnetic flux in a pore increases, the force balance demands that the outermost magnetic field lines become more nearly horizontal. When these field lines reach some critical angle, a convectively driven filamentary instability sets in^{23–25}. The resultant fluting at the boundary brings some of the outermost field lines down to the surface, where they can be grabbed and pumped downwards by granular convection. The nonlinear development of this process then leads to the formation of the penumbra, with its interlocking-comb field. This instability is subcritical, giving rise to hysteresis. As a sunspot decays, the pumping keeps most of these field lines submerged, thus explaining the existence of spots (with penumbrae) that contain less magnetic flux than the largest pores.

The submergence of some of the penumbral magnetic flux by convective pumping also explains the behaviour of the moving magnetic features (MMFs) that travel radially outward across the sunspot moat. The MMFs may be divided into three types²⁶ (Fig. 5). Type I MMFs have been interpreted as the footpoints of magnetic loops popping up from a submerged, horizontal flux sheet²⁷. As such, they fit well into our picture of flux that is held down by granular pumping, as we would expect short loops of flux to be occasionally brought to the surface by rising convective plumes. Type III MMFs would occur if the pumping mechanism occasion-

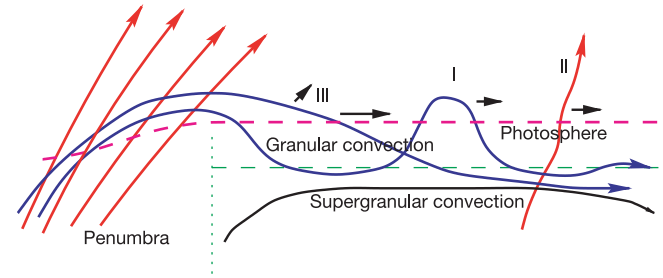


Figure 5 Moving magnetic features (MMFs) in the moat around a sunspot. Representative magnetic field lines of both the inclined magnetic field (red) in the bright filaments and the nearly horizontal magnetic field (blue) in the dark filaments are displayed. The sketch shows the configuration of MMFs of three types (I, II, III), interpreted in terms of magnetic flux pumping by granular convection in the sunspot moat. Type I MMFs are bipolar pairs of magnetic footpoints moving outward together at speeds of $0.5\text{--}1\text{ km s}^{-1}$, with the inner footpoint usually having the same polarity as the spot. Type II MMFs are single footpoints, with the same polarity as the sunspot, moving outward across the moat at speeds similar to that of type I MMFs. Type III MMFs are single footpoints with polarity opposite that of the sunspot that move rapidly outward at speeds of $2\text{--}3\text{ km s}^{-1}$.

ally allows an entire submerged penumbral flux tube to rise buoyantly and emerge through the surface at a shallow angle to the horizontal, producing a rapid outward motion of the footpoint across the moat. Type II MMFs are associated with flux tubes peeling off the sunspot and being transported horizontally outward by convection; they cause a loss of magnetic flux and decay of the sunspot. □

Received 22 May; accepted 3 October 2002; doi:10.1038/nature01174.

1. Thomas, J. H. & Weiss, N. O. (eds) *Sunspots: Theory and Observations* NATO ASI C 375 (Kluwer, Dordrecht, 1992).
2. Schmieder, B., del Toro Iniesta, J. C. & Vázquez, M. (eds) *Advances in the Physics of Sunspots* Astronomical Society of the Pacific Conference Series 118, (San Francisco, 1997).
3. Degenhardt, D. & Wiehr, E. Spatial variation of the magnetic field inclination in a sunspot penumbra. *Astron. Astrophys.* **252**, 821–826 (1991).
4. Solanki, S. K. & Montavon, C. A. P. Uncombed fields as the source of broad-band circular polarization of sunspots. *Astron. Astrophys.* **275**, 283–292 (1993).
5. Title, A. M. et al. On the magnetic and velocity field geometry of simple sunspots. *Astrophys. J.* **403**, 780–796 (1993).
6. Lites, B. W., Elmore, D. F., Seagraves, P. & Skumanich, A. Stokes profile analysis and vector magnetic fields. VI. Fine-scale structure of a sunspot. *Astrophys. J.* **418**, 928–942 (1993).
7. Solanki, S. K., Montavon, C. A. P. & Livingston, W. Infrared lines as a probe of solar magnetic features. VII. On the nature of the Evershed effect in sunspots. *Astron. Astrophys.* **283**, 221–231 (1994).
8. Rimmele, T. R. Sun center observations of the Evershed effect. *Astrophys. J.* **445**, 511–516 (1995).
9. Stanchfield, D. C. H. II, Thomas, J. H. & Lites, B. W. The vector magnetic field, Evershed flow, and intensity in a sunspot. *Astrophys. J.* **477**, 485–494 (1997).
10. Westendorp Plaza, C. et al. Evidence for a downward mass flux in the penumbral region of a sunspot. *Nature* **389**, 47–49 (1997).
11. Rüedi, I., Solanki, S. K. & Keller, C. U. Infrared lines as probes of solar magnetic features. XV. Evershed flow in cool, weak penumbral fields. *Astron. Astrophys.* **348**, L37–L40 (1999).
12. Schlichenmaier, R. & Schmidt, W. Flow geometry in a sunspot penumbra. *Astron. Astrophys.* **358**, 1122–1132 (2000).
13. Martínez Pillet, V. Spectral signature of uncombed magnetic fields. *Astron. Astrophys.* **361**, 734–742 (2000).
14. Westendorp Plaza, C., del Toro Iniesta, J. C., Ruiz Cobo, B. & Martínez Pillet, V. Optical tomography of a sunspot. III. Velocity stratification and the Evershed effect. *Astrophys. J.* **547**, 1148–1158 (2001).
15. Montesinos, B. & Thomas, J. H. The Evershed effect in sunspots as a siphon flow along a magnetic flux tube. *Nature* **390**, 485–487 (1997).
16. Stein, R. F. & Nordlund, Å. Simulations of solar granulation. I. General properties. *Astrophys. J.* **499**, 914–933 (1998).
17. Brummell, N. H., Clune, T. L. & Toomre, J. Penetration and overshooting in turbulent compressible convection. *Astrophys. J.* **570**, 825–854 (2002).
18. Tobias, S. M., Brummell, N. H., Clune, T. L. & Toomre, J. Transport and storage of magnetic field by overshooting turbulent compressible convection. *Astrophys. J.* **549**, 1183–1203 (2001).
19. Nordlund, Å. et al. Dynamo action in stratified convection with overshoot. *Astrophys. J.* **392**, 647–652 (1992).
20. Brandenburg, A. et al. Magnetic structures in a dynamo simulation. *J. Fluid Mech.* **306**, 325–352 (1996).
21. Tobias, S. M., Brummell, N. H., Clune, T. L. & Toomre, J. Pumping of magnetic fields by turbulent penetrative convection. *Astrophys. J.* **502**, L177–L180 (1998).
22. Dorc, S. B. F. & Nordlund, Å. On the transport of magnetic fields by solar-like stratified convection. *Astron. Astrophys.* **365**, 562–570 (2001).
23. Rucklidge, A. M., Schmidt, H. U. & Weiss, N. O. The abrupt development of penumbrae in sunspots. *Mon. Not. R. Astron. Soc.* **273**, 491–498 (1995).

24. Schlichenmaier, R., Jahn, K. & Schmidt, H. U. A dynamical model for the penumbral fine structure and the Evershed effect in sunspots. *Astrophys. J.* **493**, L121–L124 (1998).
25. Tildesley, M. J. On the origin of filamentary structure in sunspot penumbrae: Linear instabilities. *Mon. Not. R. Astron. Soc.* (submitted).
26. Shine, R. A. & Title, A. M. in *Encyclopedia of Astronomy and Astrophysics* 3209–3212 (Nature Publishing Group, London, and Institute of Physics Publishing, Bristol, 2001).
27. Harvey, J. & Harvey, K. Observations of moving magnetic features near sunspots. *Sol. Phys.* **28**, 61–71 (1973).

Acknowledgements We thank J. Kuwabara and Y. Uchida for assistance in producing Fig. 3. This work was supported by the UK Particle Physics and Astrophysics Research Council (J.H.T. and N.O.W.), the Sun–Earth Connection Theory programme of the US National Aeronautics and Space Administration (N.H.B. and S.M.T.), and the Nuffield Foundation (S.M.T.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.H.T. (e-mail: thomas@me.rochester.edu).

High brightness electron beam from a multi-walled carbon nanotube

Niels de Jonge*, Yann Lamy*†, Koen Schoots*‡ & Tjerk H. Oosterkamp‡

* Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands

† Ecole Supérieure de Physique et de Chimie Industrielles de la ville de Paris, 10 Rue Vauquelin, 75005 Paris, France

‡ Leiden Institute of Physics, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

Carbon nanotubes can act as electron sources¹ with very rigid structures², making them particularly interesting for use as point electron sources in high-resolution electron-beam instruments. Promising results have been reported with respect to some important requirements for such applications: a stable emitted current^{3,4,5} and a long lifetime^{6,7}. Two parameters of an electron source affect the resolution of these instruments: the energy spread of the emitted electrons and a parameter called the reduced brightness, which depends on the angular current density and the virtual source size. Several authors have measured a low energy spread associated with electron emission^{3,7,8,9}. Here we measure the reduced brightness, and find a value that is more than a factor of ten larger than provided by state-of-the-art electron sources in electron microscopes. In addition, we show that an individual multi-walled carbon nanotube emits most current into a single narrow beam. On the basis of these results, we expect that carbon nanotube electron sources will lead to a significant improvement in the performance of high-resolution electron-beam instruments.

The experiments were carried out with individual multi-walled carbon nanotubes mounted on tungsten tips. Figure 1 shows the transmission electron microscope (TEM) images of a carbon nanotube with a geometrical radius of the tube apex of 2.7 nm. As can be seen in Fig. 1b, the nanotube had a closed capping. The formation of a small probe in an electron microscope requires a small virtual source. The virtual source is the area the electrons appear to originate from when their trajectories are traced back and is usually smaller than the real emitting surface of the source¹⁰.

To determine the virtual source size, the nanotube of Fig. 1 was mounted as electron source in a point projection electron microscope in an ultrahigh vacuum system with a base pressure of 2×10^{-10} torr (ref. 11). In this setup (Fig. 2a), the emitter is positioned at a small distance (a few micrometres), z_1 , from a sharp edge provided by one of the holes in the carbon film of a TEM grid. Electrons emitted by the source generate an image on the

screen at a distance $z_2 = 16$ cm (the magnification is given by $M = z_2/z_1$). For a sufficiently small virtual source and a sufficiently sharp edge, this image shows a Fresnel interference pattern. Figure 2b shows the image of a hole in the carbon film with a clearly visible fringe pattern. A line scan across the fringes was made in a direction parallel to the long side of the rectangle drawn in the fringe pattern and plotted in Fig. 2c. A total of eight fringes was visible, using the criterion that a fringe is considered to be visible when its amplitude is not smaller than one-tenth of the amplitude of the first fringe.

The position of the n th maximum $x(n)$ with respect to the first maximum $x(0)$ in the Fresnel fringe pattern can be expressed by¹²:

$$x(n) - x(0) = z_2 \sqrt{\frac{2\lambda}{z_1}} \left(\sqrt{n + \frac{3}{8}} - \sqrt{\frac{3}{8}} \right) \quad (1)$$

where

$$x(0) = z_2 \sqrt{\frac{2\lambda}{z_1}} \frac{3}{8} \quad (2)$$

with λ the electron wavelength. The width of the fringes decreases with increasing fringe number. Equation (1) was fitted to the experimental values of the positions of the maxima as determined from the line scan (Fig. 2c) and resulted in a value of z_1 of $3.9 \mu\text{m}$ (thus the magnification $M = 4.1 \times 10^4$). The differences between the experimental values of the positions of the maxima and the positions that followed from this fit were less than 8%.

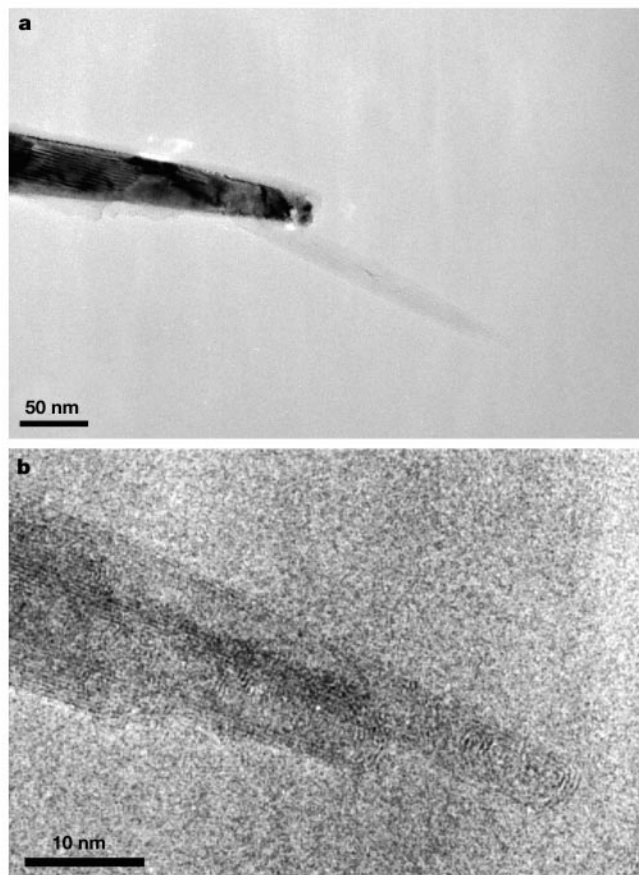


Figure 1 Transmission electron microscopy (TEM) images of an individual multi-walled carbon nanotube mounted on a tungsten tip. In **a**, the whole nanotube and the end of the tungsten tip can be seen. The glue from the carbon tape, which was used to fix the nanotube, is also visible. **b**, High-resolution TEM image of the apex of the tube with a radius of 2.7 nm.

The Fresnel fringes on a screen are convoluted with the magnified image of the source. As a consequence, the smallest visible fringe has about the same size as the magnified image of the source. The radius of the virtual source size r_v can be found from the width of the whole visible fringe pattern $x(N)$, using¹³:

$$r_v = \frac{\lambda z_2}{\pi x(N)} \quad (3)$$

where N is the number of the last visible fringe (counting from zero). For the carbon nanotube, $x(7)$ amounted to 4.0 mm. This resulted in a value of the virtual source size, $r_v = (2.1 \pm 0.2)$ nm. The validity of this method was tested by calculating the Fresnel fringe pattern for the given values of z_1 , M and λ , and convoluting the fringe pattern with a gaussian function as a model for the intensity profile of the virtual source. We did not find a deviation larger than the error margin of 10%. Similar values for the virtual source size were measured for four other samples (see Table 1).

In the upper right corner of Fig. 2b the fringe pattern of a 10-nm-wide carbon fibre is visible. The imaging of such a carbon fibre with a point projection microscope equipped with a carbon nanotube electron source was reported previously¹⁴: the authors concluded that the nanotube emitter has a high degree of coherence, as do ultrasharp tungsten sources. This implies that the source has a large spatial coherence, as well as a large temporal coherence. This is consistent with the small value of the virtual source size as measured here and the small energy spread of the carbon nanotube emitter.

Our electron source, consisting of an individual carbon nanotube, emits most of the current into a single beam; see Fig. 3. Apart from the round spot, several wide patterns were also visible, containing a minor fraction of the total current. We recorded the emission patterns of three other emitters and found a similar result. Previously, emitters containing bundles of carbon nanotubes prepared in a different way did not produce a small spot, but showed a wide emission pattern that was considered to be the magnified image of the capping of the emitting nanotube^{15,16}. The occurrence of a single bright spot is desirable for the use of this source in an electron microscope. We believe that putting a single multi-walled carbon nanotube, or a small bundle of nanotubes, on a tungsten tip

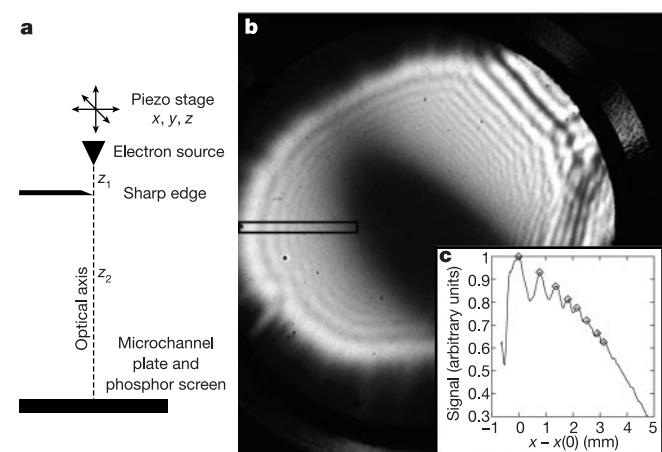


Figure 2 Measurement of the virtual source size with a point projection microscope. **a**, Schematic overview of the point projection microscope. The carbon nanotube electron source is positioned close to a sharp edge at distance z_1 with a piezo stage. A microchannel plate and phosphor screen are at a distance z_2 from the sharp edge. **b**, The Fresnel interference pattern of a hole in a thin carbon film using the same carbon nanotube as shown in Fig. 1 as electron source. This image was obtained at a voltage difference between the carbon grid and the electron source of 55 V, resulting in an emitted current of 5 nA. **c**, A line scan of integrated pixels in the rectangle drawn in the main panel, with position $x - x(0)$ relative to the position of the first maximum. The markers indicate the positions of eight maxima.

Tube	l (μ m)	r (nm)	U (V)	I (nA)	N	z_1 (μ m)	$x(N)$ (mm)	r_v (nm)
1	0.17	2.7	55	5	8	3.9	4.0	2.1
2	1.0	< 10	49	90	6	3.8	3.6	2.5
3	0.7	< 10	36	10	6	4.1	3.7	2.8
4	0.9	< 10	41	5	5	4.3	3.2	3.1
5	0.6	< 10	40	18	4	2.1	4.0	2.5

The length l and diameter r of the first tube were determined from TEM images, whereas those of the other tubes were estimated from SEM images. A measurement of the radius of the virtual source r_v was taken at the extraction voltage U , resulting in an emitted current I . The number of visible fringes is N , the distance between the tube and the carbon grid is z_1 and the width of the visible fringe pattern is indicated by $x(N)$.

is necessary to achieve a narrow beam of emitted electrons and we suspect that the result is dependent on the growth and purification method in the production of the carbon nanotubes, as well as on the method of assembly of the emitter.

The angular current density $dI/d\Omega$ of the fifth carbon nanotube sample of Table 1 was measured with a Faraday cup. A value of $dI/d\Omega = (16 \pm 3) \mu\text{A sr}^{-1}$ was found at a voltage difference between the Faraday cup and the electron source of 319 V, with a current of 14 nA in the Faraday cup and a total emitted current of 1.1 μA . This value of the angular current density is almost equal to the value of $11 \mu\text{A sr}^{-1}$ at 322 V reported elsewhere⁹.

The reduced brightness B_r is defined by:

$$B_r = \frac{dI}{d\Omega} \frac{1}{U} \frac{1}{\pi r_v^2} \quad (4)$$

where U is the acceleration voltage. For the fifth tube of Table 1 all parameters were measured. Although these measurements were taken at different conditions, we can combine the values for two reasons. First, the quantity $(dI/d\Omega)/U$ is conserved throughout an aberration-free electron optical column. Second, the virtual source size is independent of the acceleration voltage and the emitted current¹⁰. Thus, the reduced brightness of this tube amounted to $(3 \pm 1) \times 10^9 \text{ A sr}^{-1} \text{ m}^{-2} \text{ V}^{-1}$. This measured value is much larger than was expected from previous estimates of the reduced brightness of the carbon nanotube emitter^{7,9}.

The state-of-the-art electron sources for electron microscopes are the Schottky emitter¹⁷ and the cold field-emission gun¹⁸, with values of the reduced brightness varying between 10^7 and $2 \times 10^8 \text{ A sr}^{-1} \text{ m}^{-2} \text{ V}^{-1}$. Our measurements show that the carbon-nanotube electron emitter outperforms these sources by at least a factor of ten. The reduced brightness of atomically sharp emitters can be much larger than the brightness of the carbon nanotube

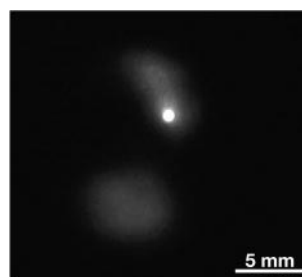


Figure 3 The emission pattern on a phosphor screen of an electron source consisting of an individual multi-walled carbon nanotube. The image was intensified by a microchannel plate positioned 1.5 cm above the emitter. The voltage difference between the emitter and the micro-channel plate was 400 V, resulting in a current of 200 nA. The electron source was transferred to a scanning electron microscope after these recordings to verify that the emitter still contained a carbon nanotube. The carbon nanotube was then removed and the emission characteristics of the clean tungsten tip were recorded for comparison. The extraction voltage required to obtain emission was much larger than with the carbon nanotube on the tip.

emitter¹³. However, these sources did not find practical application owing to their very short lifetimes. Because the energy spread of the carbon nanotube emitter^{8,9} of 0.3 eV is smaller than that of a Schottky emitter and is similar to the value of the cold field-emission gun, it is expected that the application of the carbon nanotube source in high-resolution electron-beam instruments will improve the performance of these instruments significantly. In probe-forming electron-beam instruments a trade-off has to be made between the spot size on the one hand and the electron current on the other hand. The first determines the spatial resolution, while the latter determines the time it takes to complete an image. Both aspects will benefit from the carbon nanotube source. Spatial and temporal coherence combined with a large brightness is also very important for transmission electron microscopy, energy-resolved spectroscopy and electron holography. □

Methods

An individual carbon nanotube was mounted on a tungsten tip by using a piezo nano-manipulator (Omicron) in a scanning electron microscope (SEM, Philips) as follows.

Attachment of nanotube to tungsten tip

First, a tungsten wire was welded onto a tungsten heating filament. The wire was electrochemically etched to a moderately sharp tip. This tip was transferred into the SEM and carefully pierced into carbon tape (STR tape from Shinto Paint Co.). The glue applied in this manner on the tip is necessary for a firm attachment of the nanotube.

Next, the main nanotube sample was searched for a suitable tube. The carbon nanotube sample contained multi-walled carbon nanotubes grown with the arc discharge method¹⁹. We chose a long, straight, thin tube that was uniform in diameter, pointed in the direction of the tip and was sufficiently free from other tubes to allow approaching by the tungsten tip. This tube was approached by the tip. When the tip was close enough, the tube stuck to the tip, after which it was further aligned with the axis of the tungsten tip by pulling. The breaking-off of the carbon nanotube is the most difficult part of the procedure. The carbon material is extremely strong and it can only be broken by brute force. Breaking occurs either by Joule heating with a current through the tube of more than 20 μ A, or by applying a mechanical force. Each time a nanotube emitter was made, it was tested for emission already in the SEM using a metal foil placed 5 μ m in front of the tube and a voltage difference between the tube and the foil of about 60–80 V.

Cleaning of carbon nanotube

An electron emitter prepared in this way was transferred into the pre-vacuum chamber of the ultrahigh vacuum system and 'baked' for half an hour using a halogen light bulb, after which it could enter the ultrahigh vacuum. To obtain a stable emission current, the carbon nanotube had to be cleaned carefully of molecules adsorbed on the tube and impurities in the tube structure. Therefore, the tube was first heated to a temperature of 700 °C for at least 10 min. This temperature is the carbonization temperature, at which volatile species are driven off the tube². The temperature was measured with an infrared pyrometer that was calibrated using a tungsten wire and a thermocouple inside the vacuum system. After the first heating the temperature was reduced to 600 °C and a voltage difference was applied between the emitter and the housing of the vacuum system, such that an emission current of 1 μ A was obtained. At first, the emission usually showed large fluctuations, but after about 5 min the current became more stable. Then the emission current was reduced to 100 nA and the emitter was operated for a longer period of time.

Received 8 July; accepted 16 October 2002; doi:10.1038/nature01233.

1. Saito, Y. & Uemura, S. Field emission from carbon nanotubes and its application to electron sources. *Carbon* **38**, 169–182 (2000).
2. Saito, R., Dresselhaus, G. & Dresselhaus, M. S. *Physical Properties of Carbon Nanotubes* (Imperial College Press, London, 1998).
3. Bonard, J. M., Salvétat, J. P., Stockli, T., Forro, L. & Chatelain, A. Field emission from carbon nanotubes: perspectives for applications and clues to the emission mechanism. *Appl. Phys. A* **69**, 245–254 (1999).
4. Dean, K. A., von Allmen, P. & Chalamala, B. R. Three behavioral states observed in field emission from single-walled carbon nanotubes. *J. Vac. Sci. Technol. B* **17**, 1959–1969 (1999).
5. Purcell, S. T., Vincent, P., Journet, C. & Binh, V. T. Hot nanotubes: stable heating of individual multiwall carbon nanotubes to 2000 K induced by the field-emission current. *Phys. Rev. Lett.* **88**, 105502 (2002).
6. Dean, K. A. & Chalamala, B. R. The environmental stability of field emission from single-walled carbon nanotubes. *Appl. Phys. Lett.* **75**, 3017–3019 (1999).
7. Franssen, M. J., van Rooy, T. L. & Kruit, P. Field emission energy distributions from individual multiwalled carbon nanotubes. *Appl. Surf. Sci.* **146**, 312–327 (1999).
8. Gröning, O., Kuttel, O. M., Emmenegger, C., Gröning, P. & Schlappbach, L. Field emission properties of carbon nanotubes. *J. Vac. Sci. Technol. B* **18**, 665–678 (2000).
9. de Jonge, N. & van Druten, N. J. Field emission from individual multiwalled carbon nanotubes prepared in an electron microscope. *Ultramicroscopy* (in the press).
10. Hawkes, P. W. & Kasper, E. *Principles of Electron Optics II: Applied Geometrical Optics* (Academic, London, 1996).
11. Franssen, M. J. et al. Characterization of ultrasharp field emitters by projection microscopy. *Appl. Surf. Sci.* **94/95**, 107–112 (1995).
12. Hawkes, P. W. & Kasper, E. *Principles of Electron Optics III: Wave Optics* (Academic, London, 1996).

13. Spence, J. C. H., Qian, W. & Silverman, M. P. Electron source brightness and degeneracy from Fresnel fringes in field emission point projection microscopy. *J. Vac. Sci. Technol. A* **12**, 542–547 (1994).
14. Schmid, H. & Fink, H. W. Carbon nanotubes are coherent electron sources. *Appl. Phys. Lett.* **70**, 2679–2680 (1997).
15. Dean, K. A. & Chalamala, B. R. Field emission microscopy of carbon nanotube caps. *J. Appl. Phys.* **85**, 3832–3836 (1999).
16. Saito, Y. et al. Conical beams from open nanotubes. *Nature* **389**, 554–555 (1997).
17. van Veen, A. H. V., Hagen, C. W., Barth, J. E. & Kruit, P. Reduced brightness of the ZrO/W Schottky electron emitter. *J. Vac. Sci. Technol. B* **19**, 2038–2044 (2001).
18. Orloff, J. Survey of electron sources for high-resolution microscopy. *Ultramicroscopy* **28**, 88–97 (1989).
19. Colbert, D. T. et al. Growth and sintering of fullerene nanotubes. *Science* **266**, 1218–1222 (1994).

Acknowledgements We thank M. Kaiser for the TEM images, N. J. van Druten and B. Buijsse for discussions, T. van Rooij for help with the vacuum system and R. E. Smalley for providing the carbon nanotube sample. N.d.J. was supported by FEI company. T.H.O. was supported by Nederlandse organisatie voor wetenschappelijk onderzoek and stichting voor fundamenteel onderzoek der materie.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.J. (e-mail: niels.de.jonge@philips.com).

Exchange-coupled nanocomposite magnets by nanoparticle self-assembly

Hao Zeng^{*†}, Jing Li[‡], J. P. Liu[†], Zhong L. Wang[‡] & Shouheng Sun^{*}

^{*} IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA
[†] Institute for Micromanufacturing, Louisiana Tech University, Ruston, Louisiana 71272, USA
[‡] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Exchange-spring magnets are nanocomposites that are composed of magnetically hard and soft phases that interact by magnetic exchange coupling¹. Such systems are promising for advanced permanent magnetic applications, as they have a large energy product—the combination of permanent magnet field and magnetization—compared to traditional, single-phase materials^{1–3}. Conventional techniques, including melt-spinning^{4–6}, mechanical milling^{7–9} and sputtering^{10–12}, have been explored to prepare exchange-spring magnets. However, the requirement that both the hard and soft phases are controlled at the nanometre scale, to ensure efficient exchange coupling, has posed significant preparation challenges. Here we report the fabrication of exchange-coupled nanocomposites using nanoparticle self-assembly. In this approach, both FePt and Fe₃O₄ particles are incorporated as nanometre-scale building blocks into binary assemblies. Subsequent annealing converts the assembly into FePt–Fe₃Pt nanocomposites, where FePt is a magnetically hard phase and Fe₃Pt a soft phase. An optimum exchange coupling, and therefore an optimum energy product, can be obtained by independently tuning the size and composition of the individual building blocks. We have produced exchange-coupled isotropic FePt–Fe₃Pt nanocomposites with an energy product of 20.1 MG Oe, which exceeds the theoretical limit of 13 MG Oe for non-exchange-coupled isotropic FePt by over 50 per cent.

Hexane dispersions of FePt (refs 13, 14) and Fe₃O₄ (ref. 15) nanoparticles with selected concentration, volume and sizes were mixed under ultrasonic agitation, and three-dimensional binary assemblies were induced by either evaporation of the hexane or addition of ethanol. The mass ratio of Fe₃O₄ and FePt was controlled in the range from 1:5 to 1:20, while the size of FePt

was kept at 4 nm and Fe_3O_4 was varied from 4 nm to 12 nm. Figure 1 shows transmission electron microscopy (TEM) images of Fe_3O_4 : $\text{Fe}_{58}\text{Pt}_{42}$ binary assemblies with different sizes and fixed mass ratio of 1:10. Depending on the sizes of nanoparticles, we observed three different assembly structures. For the 4 nm:4 nm assembly (Fig. 1a), $\text{Fe}_{58}\text{Pt}_{42}$ and Fe_3O_4 randomly occupy the sites in a hexagonal lattice; but for the 8 nm:4 nm assembly (Fig. 1b), a typical local ordering appears with each big particle (Fe_3O_4) surrounded by 6–8 small particles ($\text{Fe}_{58}\text{Pt}_{42}$). A large difference in particle sizes in an assembly, such as in a 12 nm:4 nm assembly (Fig. 1c), however, results in clear phase segregation with the 12 nm and 4 nm particles forming their own particle lattice arrays. These ordering structures depend mainly on the particle size ratio, as in the case of Au particle

assembly¹⁶, and are insensitive to the mass ratio range we studied.

The binary assemblies were converted into FePt– Fe_3Pt nanocomposites by annealing under a flow of Ar + 5% H_2 at 650 °C for 1 h. The annealing reduces iron oxide to iron¹⁵, and transforms the FePt from disordered face-centred cubic (f.c.c.) to ordered face-centred tetragonal (f.c.t.) structure¹³ that possesses high magnetocrystalline anisotropy ($> 5 \times 10^7 \text{ erg cm}^{-3}$) (refs 17, 18), providing a large coercivity H_c in the nanocomposites. The high temperature treatment under gas flow also desorbs the organic stabilizers around each

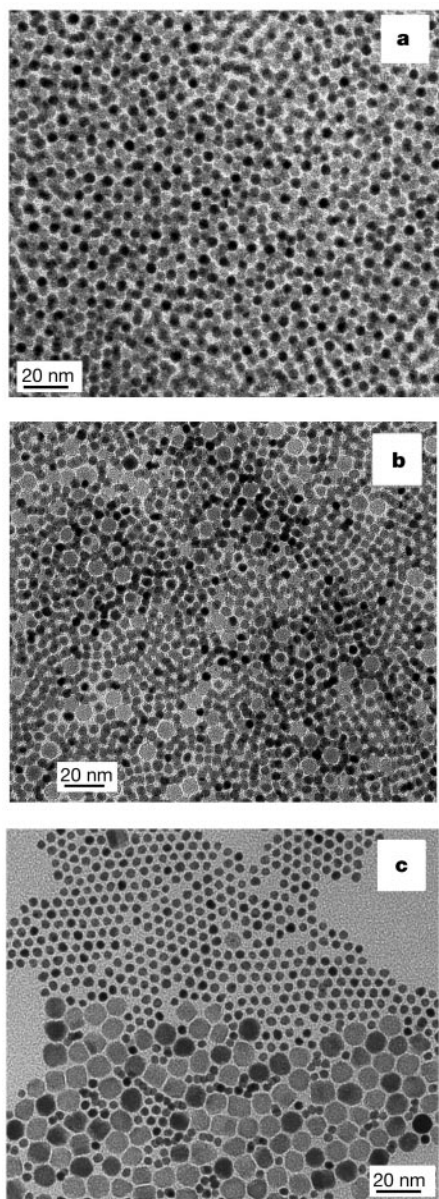


Figure 1 TEM images showing binary nanoparticle assemblies. **a**, Fe_3O_4 (4 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly; **b**, Fe_3O_4 (8 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly; and **c**, Fe_3O_4 (12 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly. The assembly contained Fe_3O_4 and FePt binary nanoparticles with a mass ratio of Fe_3O_4 :FePt = 1/10 and was formed by solvent evaporation of the mixed nanoparticle dispersions on amorphous carbon-coated TEM grids. All images were acquired using a Philips CM12 microscope at 120 kV.

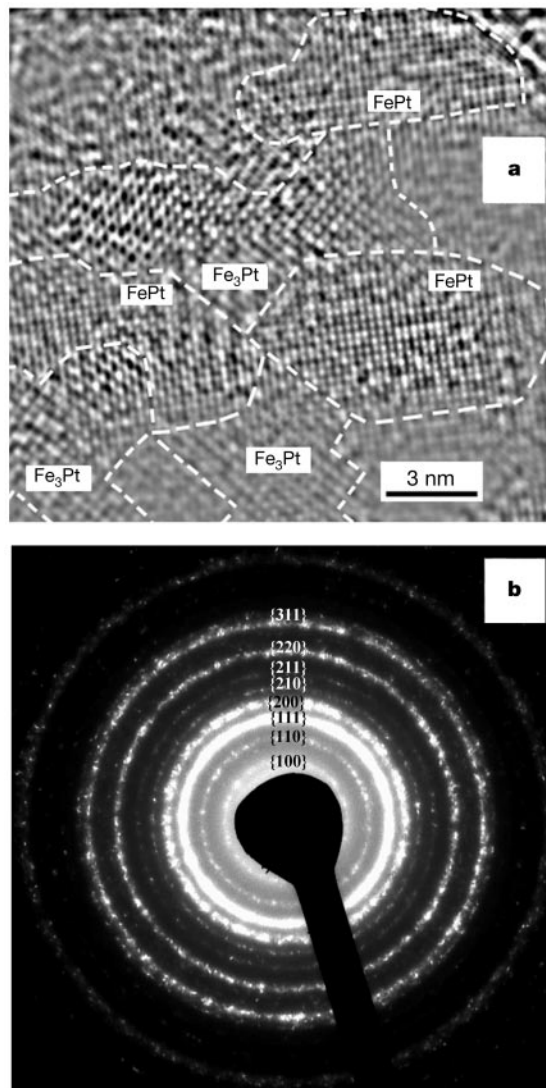


Figure 2 Structural analyses of the FePt– Fe_3Pt nanocomposite obtained from the annealed Fe_3O_4 (4 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly. **a**, A typical HRTEM image for a sintered FePt– Fe_3Pt particle. The FePt and Fe_3Pt phases are in coexistence as different domains within the particle, with each domain having a dimension of about 5 nm, showing a modulated FePt– Fe_3Pt spatial distribution. Here, the FePt (ordered f.c.t. structure) and Fe_3Pt (ordered f.c.c. structure) phases were identified by their different [001] projected potential, owing to their different composition modulation periodicities. The image was acquired using a Jeol 4000EX HRTEM at 400 kV. To prepare the TEM samples, the composite was deposited on a solid NaCl substrate. The substrate was then removed in water, and the composite was thinned using ion milling for TEM observation. **b**, A typical SAD pattern of the FePt– Fe_3Pt nanocomposite. The f.c.t.-structured FePt was indexed in the ring pattern. The diffraction rings of the f.c.c.-structured Fe_3Pt cannot be resolved from those of the FePt phase. The ring pattern of the diffraction demonstrates that the nanocrystallites in the nanocomposite are three-dimensionally randomly oriented.

particle, allowing the nanoparticles to sinter. The partial interdiffusion between Fe and FePt creates a new f.c.c.-structured phase, Fe₃Pt, which is magnetically soft with high magnetization¹⁹. Figure 2a is a typical high-resolution TEM (HRTEM) image of a sintered sample obtained from a 4 nm:4 nm assembly. Structural analysis reveals that the coalesced particle is divided into two distinct phases with dimensions of the order of 5 nm: a f.c.c. FePt phase that is magnetically hard, and a f.c.c. Fe₃Pt phase that is magnetically soft. Elemental analysis from spatially resolved energy dispersive spectroscopy (EDS) confirms the existence of two phases with the compositions of Fe:Pt close to 1:1 and 3:1 respectively. After examining various annealed 4 nm:4 nm and 8 nm:4 nm samples with different initial mass ratios, we found that: (1) the Fe₃Pt phase is uniformly dispersed into the FePt matrix; and (2) its dimension is below 10 nm. Figure 2b is a typical selected area diffraction (SAD) pattern of a FePt–Fe₃Pt nanocomposite. The ring pattern in SAD indicates three-dimensional random crystal orientation. The SAD patterns from different tilting angles show similar ring patterns, demonstrating that the isotropic FePt–Fe₃Pt nanocomposite has formed. Trace amounts of α -Fe probably exist in the composite from the 8 nm:4 nm assembly, as suggested by both TEM and EDS analyses, whereas in the annealed 12 nm:4 nm sample, we observe large α -Fe particles over 20 nm in diameter, due to phase segregation observed in the as-prepared sample.

Magnetic properties of FePt–Fe₃Pt nanocomposites vary with different initial mass ratios of Fe₃O₄ and Fe₅₈Pt₄₂. Figure 3 shows the saturation magnetization (M_s), remanent magnetization (M_r) and coercivity (H_c) of the FePt–Fe₃Pt composites from the Fe₃O₄ (4 nm):Fe₅₈Pt₄₂ (4 nm) assemblies as a function of the initial mass ratio, Fe₃O₄:Fe₅₈Pt₄₂. It can be seen that M_s increases monotonically from 950 e.m.u. cm⁻³ for pure FePt to 1,110 e.m.u. cm⁻³ for the 1:5 mass ratio, while H_c decreases sharply from 19 kOe to 6.8 kOe. The M_s of the pure FePt assembly is about 15% lower than the 1,100 e.m.u. cm⁻³ reported for bulk FePt¹⁸. This could result from nanoparticle size, surface and composition effects. M_r shows a maximum of 740 e.m.u. cm⁻³ at the 1:10 mass ratio, a 17% increase from pure FePt. The remanence ratio M_r/M_s values for all samples are greater than 0.6. Figure 4a is a representative hysteresis loop for the composite from the 4 nm:4 nm assembly with 1:10 mass ratio. Although the sample consists of both magnetically hard and soft

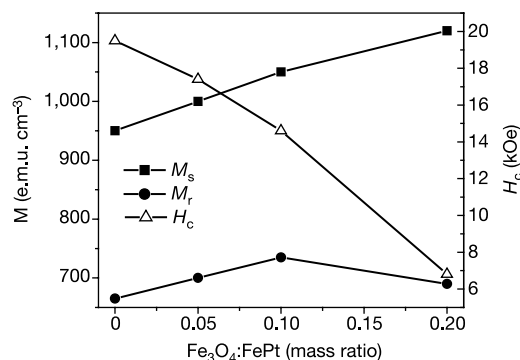


Figure 3 M_s , M_r and H_c of the FePt–Fe₃Pt nanocomposites from the annealed Fe₃O₄ (4 nm):Fe₅₈Pt₄₂ (4 nm) assemblies as a function of mass ratio of Fe₃O₄:FePt. All the data were collected at room temperature on a superconducting quantum interference device (SQUID) magnetometer with field up to 7 T. The magnetic moment is measured as σ (moment per unit mass), and converted to M (moment per unit volume) assuming an ideal density derived from the final Fe and Pt composition of the composite. The elemental composition of the nanocomposite was obtained by inductively coupled plasma–optical emission spectrometry. The impurity level is less than 2%. Because the complete saturation is not reached under a field of up to 7 T, M measured at 7 T is used to represent the lower bound of M_s .

phases, the hysteresis measurements show that the magnetization changes smoothly with field, similar to the hysteresis behaviour of a single-phase material²⁰, indicating that the exchange coupling between the two phases has been realized. The hysteresis measurements also reveal that the remanence of the two-phase nanocomposites is higher than that of the single-phase FePt, which is further evidence of effective hard–soft exchange coupling.

The magnetic behaviour of the FePt–Fe₃Pt nanocomposites changes significantly with the initial sizes of Fe₃O₄ and Fe₅₈Pt₄₂. For the composite from Fe₃O₄ (8 nm):Fe₅₈Pt₄₂ (4 nm) assembly with a mass ratio of 1:10, hysteresis measurements show a single-phase-like loop (not shown here) with large H_c at about 24 kOe and enhanced remanence; but for the composite from Fe₃O₄ (12 nm):Fe₅₈Pt₄₂ (4 nm) assembly with the same mass ratio, the loop shows a kink at low field (Fig. 4b). These changes in hysteresis behaviour correlate directly to size-dependent assembly structures observed in the TEM images (Fig. 1). The Fe₃O₄ (8 nm):Fe₅₈Pt₄₂ (4 nm) assembly possesses a well-mixed structure, with all Fe₃O₄ being surrounded by Fe₅₈Pt₄₂ particles (Fig. 1b). Such a structure leads to a modulated Fe₃Pt soft phase and FePt hard phases in the annealed composite with the soft grains limited to sizes less than 10 nm, as confirmed by HRTEM analyses. For the composite from a Fe₃O₄ (12 nm):Fe₅₈Pt₄₂ (4 nm) assembly, however, due to the spatial segregation of Fe₅₈Pt₄₂ and Fe₃O₄ particles (Fig. 1c), large Fe particles over 20 nm in diameter are formed. For effective exchange coupling to occur within a two-phase magnet, the dimension of the soft phase should be smaller than twice the domain wall width of the hard phase, at about 10 nm (refs 1–3). Therefore, in a composite with a large soft phase, hard and soft phases are not able to switch cooperatively. As a result, its hysteresis loop shows a two-phase behaviour (Fig. 4b).

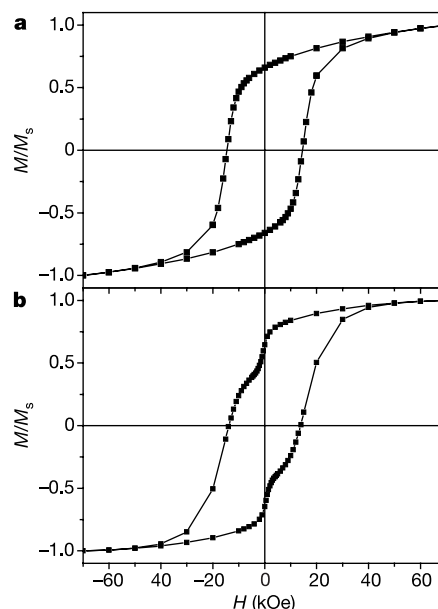


Figure 4 Typical hysteresis loops of two FePt-based nanocomposites. The composites were made from the annealed Fe₃O₄:FePt assemblies with particle mass ratio being kept constantly at Fe₃O₄:FePt = 1:10. **a**, FePt–Fe₃Pt nanocomposite from Fe₃O₄ (4 nm):Fe₅₈Pt₄₂ (4 nm) assembly. The loop shows single-phase-like behaviour, indicating effective exchange coupling between FePt and Fe₃Pt. **b**, A nanocomposite from Fe₃O₄ (12 nm):Fe₅₈Pt₄₂ (4 nm) assembly. Owing to the phase separation in the 12 nm:4 nm assembly as illustrated in Fig. 1c, the annealed sample contained large body centred cubic (b.c.c.) Fe grains as confirmed by HRTEM and EDS, rendering a nanocomposite with the hysteresis showing two-phase behaviour. The kink at low field is related to the magnetization reversal of the soft Fe.

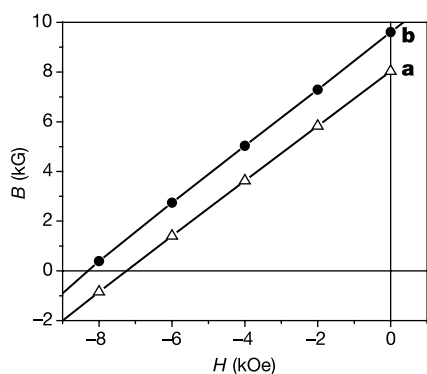


Figure 5 Second-quadrant B - H curves for the annealed samples. **a**, An annealed 4 nm $\text{Fe}_{58}\text{Pt}_{42}$ nanoparticle assembly. **b**, A hard-soft exchange-coupled FePt - Fe_3Pt nanocomposite. The composite was obtained from the annealed Fe_3O_4 (4 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly (mass ratio $\text{Fe}_3\text{O}_4:\text{FePt} = 1:10$). The energy product, $(BH)_{\text{max}}$, describes the available energy density of the materials and is defined as the maximum BH product of the second-quadrant B - H curve²¹.

The optimized nanostructures for exchange coupling yield both high remanent magnetization and coercivity, resulting in an enhanced energy product $(BH)_{\text{max}}$ (ref. 21). Figure 5 illustrates B - H curves of an annealed 4 nm $\text{Fe}_{58}\text{Pt}_{42}$ nanoparticle assembly (Fig. 5a), and a nanocomposite from a Fe_3O_4 (4 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly with a 1:10 mass ratio (Fig. 5b). The measured $(BH)_{\text{max}}$ for the single-phase $\text{Fe}_{58}\text{Pt}_{42}$ is 14.7 MG Oe. For the nanocomposite derived from Fe_3O_4 (4 nm): $\text{Fe}_{58}\text{Pt}_{42}$ (4 nm) assembly, $(BH)_{\text{max}}$ reaches 20.1 MG Oe, exceeding the value for single phase $\text{Fe}_{58}\text{Pt}_{42}$ assembly by 37%, and the theoretical limit of 13 MG Oe for non-exchange-coupled isotropic FePt by over 50%. This $(BH)_{\text{max}}$ enhancement clearly indicates effective exchange coupling between the hard and soft phases.

We have described the fabrication of exchange-coupled isotropic FePt - Fe_3Pt nanocomposite magnets by self-assembly of binary FePt and Fe_3O_4 nanoparticles, and controlled annealing. By engineering the nanoscale dimension and spatial distribution of the hard and soft phases, an enhanced energy product is achieved. Our approach shows great potential for future fabrication of high-performance exchange-spring magnets. Further, it can be extended to other multi-component systems, providing both a model for studying fundamental relationships between nanostructure and interparticle interactions, and a practical route to functional nanocomposites and devices. □

Received 15 July; accepted 10 October 2002; doi:10.1038/nature01208.

- Kneller, E. F. & Hawig, R. The exchange-spring magnet: a new material principle for permanent magnets. *IEEE Trans. Magn.* **27**, 3588–3600 (1991).
- Skomski, R. & Coey, J. M. D. Giant energy product in nanostructured two-phase magnets. *Phys. Rev. B* **48**, 15812–15816 (1993).
- Schrefl, T., Kronmüller, H. & Fidler, J. Exchange hardening in nano-structured two-phase permanent magnets. *J. Magn. Mag. Mater.* **127**, L273–L277 (1993).
- Coehoorn, R., De Mooij, D. B. & De Waard, C. Meltsun permanent magnet materials containing Fe_3B as the main phase. *J. Magn. Mag. Mater.* **80**, 101–104 (1989).
- Ping, D. H., Hono, K. & Hirotsawa, S. Partitioning of Ga and Co atoms in a $\text{Fe}_3\text{B}/\text{Nd}_2\text{Fe}_{14}\text{B}$ nanocomposite magnet. *J. Appl. Phys.* **83**, 7769–7775 (1998).
- Kobayashi, T., Yamasaki, M. & Hamano, M. Mössbauer study on intergranular phases in the bcc-Fe/NdFeB nanocomposite alloys. *J. Appl. Phys.* **87**, 6579–6581 (2000).
- Gong, W., Hadjippanayis, G. C. & Krause, R. F. Mechanically alloyed nanocomposite magnets. *J. Appl. Phys.* **75**, 6649–6651 (1994).
- McCormick, P. G. et al. Mechanically alloyed nanocomposite magnets. *J. Appl. Phys.* **83**, 6256–6261 (1998).
- Zhang, J., Zhang, S.-Y., Zhang, H.-W., Shen, B.-G. & Li, B.-H. Structure and magnetic properties of $\text{Sm}_2\text{Co}_5/\alpha\text{-Fe}$ ($x = 0.65$ – 1.3) prepared by mechanical milling and subsequent annealing. *J. Appl. Phys.* **89**, 2857–2860 (2001).
- Fullerton, E. E., Jiang, J. S., Sowers, C. H., Pearson, J. E. & Bader, S. D. Structure and magnetic properties of exchange-spring Sm-Co/Co superlattices. *Appl. Phys. Lett.* **72**, 380–382 (1997).
- Liu, J. P., Luo, C. P., Liu, Y. & Sellmyer, D. J. High energy products in rapidly annealed nanoscale Fe/Pt multilayers. *Appl. Phys. Lett.* **72**, 483–485 (1998).

- Liu, J. P., Liu, Y., Skomski, R. & Sellmyer, D. J. Magnetic hardening in SmCo_5 -Co multilayers and nanocomposites. *J. Appl. Phys.* **85**, 4812–4814 (1999).
- Sun, S., Murray, C. B., Weller, D., Folks, L. & Moser, A. Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. *Science* **287**, 1989–1992 (2000).
- Sun, S., Fullerton, E. E., Weller, D. & Murray, C. B. Compositionally controlled FePt nanoparticle materials. *IEEE Trans. Magn.* **37**, 1239–1243 (2001).
- Sun, S. & Zeng, H. Size controlled synthesis of magnetite nanoparticles. *J. Am. Chem. Soc.* **124**, 8204–8205 (2002).
- Kiely, C. J., Fink, J., Brust, M., Bethell, D. & Schiffrin, D. J. Spontaneous ordering of bimodal ensembles of nanoscopic gold clusters. *Nature* **396**, 444–446 (1998).
- Weller, D. et al. High K_u materials approach to 100 Gbits/in². *IEEE Trans. Magn.* **36**, 10–15 (2000).
- Klemmer, T., Hoydick, D., Okumura, H., Zhang, B. & Soffa, W. A. Magnetic hardening and coercivity mechanisms in L_{10} ordered FePd ferromagnets. *Scripta Met.* **33**, 1793–1805 (1995).
- Podgorny, M. Electronic structure of the ordered phases of Pt-Fe alloys. *Phys. Rev. B* **43**, 13700–13718 (1991).
- Zeng, H. et al. Exchange-coupled FePt nanoparticle assembly. *Appl. Phys. Lett.* **80**, 2583–2585 (2002).
- O'Handley, R. C. *Modern Magnetic Materials* 469–472 (John Wiley & Sons, New York, 2000).

Acknowledgements This work is supported in part by the US Defense Advanced Research Program Agency (DARPA) through the Army Research Office (ARO). H. Z. and J. L. thank DARPA for support through Louisiana Tech University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.S. (e-mail: ssun@us.ibm.com).

Active transport of Ca^{2+} by an artificial photosynthetic membrane

Ira M. Bennett*, Hebe M. Vanegas Farfano*, Federica Bogani*, Alex Primak*, Paul A. Liddell*, Luis Otero†, Leonides Sereno†, Juana J. Silber†, Ana L. Moore*, Thomas A. Moore* & Devens Gust*

* Department of Chemistry and Biochemistry and Center for the Study of Early Events in Photosynthesis, Arizona State University, Tempe, Arizona 85287, USA

† Departamento de Química y Física, Universidad Nacional de Río Cuarto, Río Cuarto, Argentina

Transport of calcium ions across membranes and against a thermodynamic gradient is essential to many biological processes, including muscle contraction, the citric acid cycle, glycogen metabolism, release of neurotransmitters, vision, biological signal transduction and immune response. Synthetic systems that transport metal ions across lipid or liquid membranes are well known^{1–6}, and in some cases light has been used to facilitate transport⁷. Typically, a carrier molecule located in a symmetric membrane binds the ion from aqueous solution on one side and releases it on the other. The thermodynamic driving force is provided by an ion concentration difference between the two aqueous solutions, coupling to such a gradient in an auxiliary species, or photomodulation of the carrier by an asymmetric photon flux⁷. Here we report a different approach, in which active transport is driven not by concentration gradients, but by light-induced electron transfer in a photoactive molecule that is asymmetrically disposed across a lipid bilayer. The system comprises a synthetic, light-driven transmembrane Ca^{2+} pump based on a redox-sensitive, lipophilic Ca^{2+} -binding shuttle molecule whose function is powered by an intramembrane artificial photosynthetic reaction centre. The resulting structure transports calcium ions across the bilayer of a liposome to develop both a calcium ion concentration gradient and a membrane potential, expanding Mitchell's concept of a redox loop mechanism for protons⁸ to include divalent cations. Although the quantum yield is relatively low (~ 1 per cent), the Ca^{2+} electrochemical potential developed is significant.

A redox loop mechanism for transmembrane ion transport is premised on differential binding of an ion by reduced and oxidized forms of a lipid-soluble transporter (shuttle) molecule. The transmembrane Ca^{2+} transport system is based on a redox loop involving the synthetic, lipophilic, membrane-soluble shuttle **1** (Fig. 1). Inspiration for **1** came from a suggestion that transmembrane Ca^{2+} transport in sponges might be based on a hydroquinone Ca^{2+} chelator⁹. Calcium binding by the hydroquinone form of **1** is expected, based on results for related 9,10-anthraquinone and 9-anthrone derivatives⁶. This conclusion is supported by experiments on 2,5-dihydroxyacetophenone which, unlike **1**, has solubility in polar solvents. Calcium ion binding by this compound in aqueous solution at pH = 7 was investigated by using a calcium-specific electrode. Binding by Ca^{2+} of one or two hydroquinone anions with equilibrium constants of 820 M^{-1} and 75 M^{-1} , respectively, was detected. Binding was also studied using nuclear magnetic resonance (NMR) spectroscopy. Addition of excess CaCl_2 in D_2O to 2,5-dihydroxyacetophenone in deuterated acetonitrile containing one equivalent of KOH led to downfield shifts of the aromatic proton resonances, consistent with the binding of Ca^{2+} . A similar experiment with the quinone form resulted in no detectable shifts. We conclude that lipid-soluble hydroquinone **1** binds calcium ions, whereas the corresponding quinone does not.

We incorporated **1** in its hydroquinone form into the phospholipid bilayer of a liposome so as to serve as one component of a Ca^{2+}

pump. The proposed mechanism suggested by Fig. 1 stipulates that during a pumping cycle, a minimum of two molecules of the hydroquinone at the external surface of the membrane lose protons to the aqueous phase (pH 7.5) and bind Ca^{2+} . The resulting lipid-soluble complex diffuses across the membrane to the opposite interface. Oxidation of hydroquinone generates quinone, releasing Ca^{2+} to the aqueous phase inside the liposome. Diffusion of quinone and neutral hydroquinone (carrying two protons) to the initial side of the membrane, followed by reduction of the quinone, completes the redox loop. The result is cyclic transport of Ca^{2+} with no concurrent transport of hydrogen ions.

The redox power for the pump comes from light through artificial photosynthetic reaction centre **2**, which was employed in a previously reported light-powered transmembrane proton pump^{10,11}. This carotenoid (C), porphyrin (P), naphthoquinone (Q), molecular triad was inserted in a vectorial manner into the liposomal membrane with its naphthoquinone end near the external surface and its carotenoid functionality inside the hydrophobic portion of the bilayer¹¹. Excitation of **2** by visible light produces the $\text{C}^+-\text{P}-\text{Q}^-$ charge-separated state through photoinduced electron transfer. The carotenoid radical cation, a strong oxidizing agent, is responsible for conversion of the calcium-binding hydroquinone form of **1** to the quinone (either by sequential one-electron oxidations or by disproportionation reactions of the semiquinone intermediate), with release of Ca^{2+} to the inner liposomal volume. The naphthoquinone radical anion, located at the external surface of the liposome, is available for reduction of the quinone form of **1**.

Irradiation of the assembly diagrammed in Fig. 1 with light absorbed by the porphyrin leads to transmembrane Ca^{2+} transport, as indicated in Fig. 2. The internal Ca^{2+} concentration was monitored using the fluorescent dye Fluo-3 (ref. 12), which was located only inside the liposomes, and whose response was calibrated using liposomes containing known calcium concentrations. Figure 2 also shows a control experiment without illumination. When triad **2** or shuttle **1** was omitted, no calcium transport was detected. Calcium transport was not observed when **1** was replaced with a redox-sensitive proton shuttle that forms a transmembrane proton

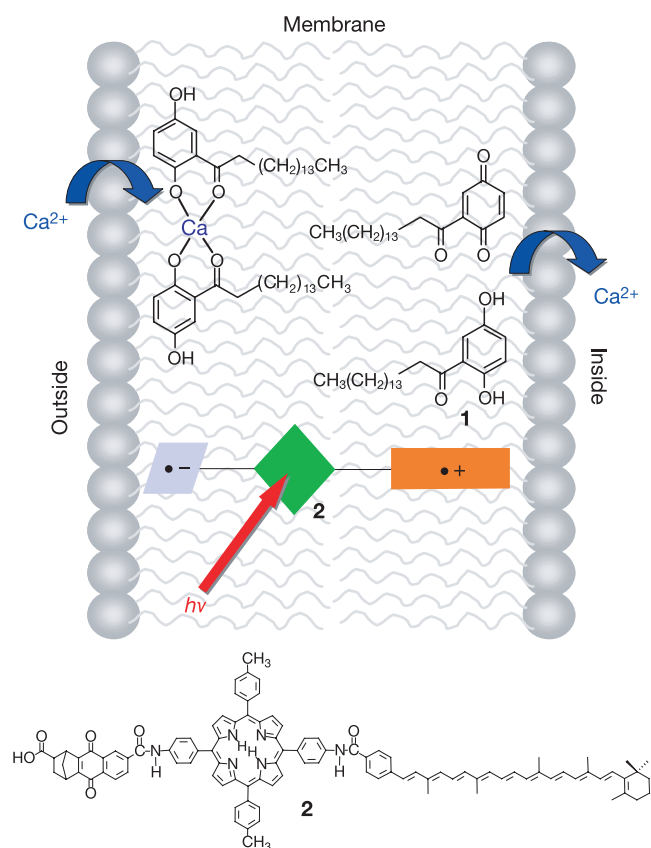


Figure 1 Schematic representation of a liposome-based, light-powered transmembrane Ca^{2+} pump. The carotenoid–porphyrin–naphthoquinone artificial reaction centre **2** crosses the membrane, with its hydrophilic quinone moiety near the external interface. Shuttle molecule **1** chelates Ca^{2+} at the external interface, diffuses across the membrane, and is oxidized by the carotenoid radical cation formed by photoinduced electron transfer in molecule **2**. After release of Ca^{2+} to the internal aqueous volume, the quinone form of **1** diffuses back across the membrane, where it can be reduced by the radical anion of **2** formed by photoinduced electron transfer.

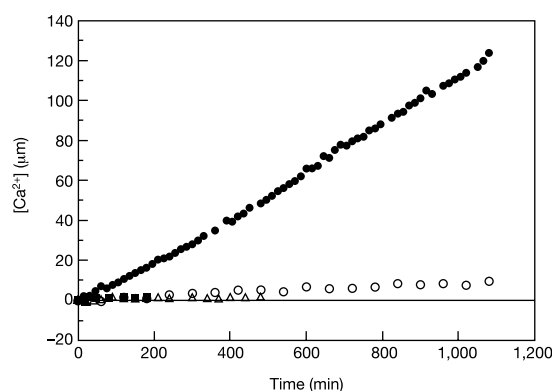


Figure 2 Transmembrane Ca^{2+} transport. The concentration of Ca^{2+} transported into the liposome interior as a function of time with (filled circles) and without (open circles) irradiation with 633-nm light was determined from the calibrated fluorescence intensity of the calcium indicator Fluo-3. Also shown are results for similar samples lacking the shuttle **1** (filled squares), and in which **1** has been replaced by a similar molecule bearing a methoxy group in the 5-position (open triangles). Reverse-phase evaporation liposomes prepared as previously described¹¹ were used. The organic lipid-forming solution contained 0.4 mol% of molecule **1**, and the aqueous solution (pH 7.5) contained 150 mM KCl, 10 mM HEPES buffer, and $100 \mu\text{M}$ Fluo-3. Purification of the liposomes and vectorial insertion of **2** were carried out as previously described¹¹, and CaCl_2 was then added to 1.0 mM. Actinic light was provided by a 3.6 mW, about 1 cm beam of 633-nm laser light. The sample in the cuvette (1.5 ml) was stirred.

pump¹⁰. Use of the ΔpH -sensitive dye 9-aminoacridine demonstrated that at $\text{pH} = 7.5$, in the presence or absence of added Ca^{2+} , **1** does not transport hydrogen ions across the membrane, with or without irradiation, under the conditions used for Ca^{2+} pumping. Methylation of the oxygen at the 5-position of **1** precludes formation of the quinone, increases the first oxidation potential (0.57 V versus 0.20 V (standard calomel electrode, SCE) for **1**) and abolishes Ca^{2+} pumping.

The suggested mechanism requires Ca^{2+} pumping to be electrogenic, as there is no net transmembrane transport of other charges. Introduction and calibration of the dye 8-anilino-1-naphthalenesulphonic acid, whose fluorescence yield is a function of membrane potential ($\Delta\psi$), and elimination of Fluo-3 allowed us to investigate this possibility. Illumination under conditions for Ca^{2+} pumping leads to formation of a membrane potential that is positive on the interior of the liposomes (Fig. 3). No $\Delta\psi$ is produced without actinic light, without **1** or **2**, or when **1** is replaced by its 5-methoxy analogue. Without added Ca^{2+} , and with the addition of a strong calcium chelator such as EDTA to the external aqueous solution to bind any adventitious calcium ions, a membrane potential is not produced. The membrane potential is discharged upon addition of 50 μM valinomycin, which permits potassium ions to flow out of the liposomes, or eventually upon standing in the dark.

In Fig. 2 the calcium concentration inside the liposome increases approximately linearly with irradiation time up to about 1,000 min. Because of the large binding constant of Fluo-3 for Ca^{2+} ($K_d = 390 \text{ nM}$ at $\text{pH} = 7.2$), there is never any significant amount of free Ca^{2+} inside the liposome, and therefore no free Ca^{2+} concentration opposing Ca^{2+} transport. In the absence of the Ca^{2+} -binding dye, however, the membrane potential studies may be used to obtain quantitative information about the performance of the Ca^{2+} pump. Figure 3 demonstrates that when no Ca^{2+} is added to the liposome interior before illumination, $\Delta\psi$ builds up with Ca^{2+} pumping, and maximizes at about 120 mV. At this point, the Ca^{2+} concentration inside the liposomes is still about 20 times lower than on the outside (Fig. 2). Even though the ion pump has not reversed the Ca^{2+} concentration gradient, potential energy is being stored as a membrane potential of about 120 mV. The fact that in the presence of Fluo-3, Ca^{2+} pumping continues unabated even after this potential limit is reached (Fig. 2) suggests that beyond this

limit, charge import by the Ca^{2+} pump is balanced by leakage of positive ions, or negative ions such as chloride.

Figure 3 also shows the buildup of $\Delta\psi$ when the initial concentration of Ca^{2+} both inside and outside the liposomes is 1.0 mM. The increase of $\Delta\psi$ with illumination occurs essentially as it does when calcium is excluded from the interior, and no $\Delta\psi$ is generated in the absence of actinic light. Thus Ca^{2+} can be pumped across the membrane even against a Ca^{2+} concentration gradient and a membrane potential. Because the amount of Ca^{2+} pumped ($\sim 50 \mu\text{M}$) is small relative to the ambient concentration, the electrochemical potential built up is mainly $\Delta\psi$. This experiment unambiguously demonstrates that the transmembrane ion pump stores some of the light energy harvested by the artificial reaction centre as electrical and chemical potential energy.

The $\Delta\psi$ measurements may be used to estimate the quantum yield of Ca^{2+} transport, $\Phi_{\text{Ca}^{2+}}$, assuming that the only electrogenic process is the import of two positive charges per ion. The initial rate of $\Delta\psi$ increase ($d\Delta\psi_{\text{Ca}^{2+}}/dt$) was compared with the initial rate of $\Delta\psi$ increase in an artificial proton pump/ATP producing system ($d\Delta\psi_{\text{H}^+}/dt$) described earlier. The quantum yield of the proton pump was previously estimated¹¹ as 0.6. The actinic light level was such that the rate of $\Delta\psi$ generation was linear with intensity for both pumps. Under these conditions, $(d\Delta\psi_{\text{H}^+}/dt)/(d\Delta\psi_{\text{Ca}^{2+}}/dt)$ is proportional to the ratio of the quantum yields, giving $\Phi_{\text{Ca}^{2+}} \approx 0.01$. This method does not require knowledge of the size or number of liposomes, as long as these quantities and the number of artificial reaction centres per liposome are the same in both cases.

The results reported here are consistent with the proposed mechanism for Ca^{2+} translocation suggested above. However, the details of the transport are not as yet fully understood. In particular, the stoichiometry of Ca^{2+} binding, the exact oxidation state of **1** at the time of Ca^{2+} release, and the detailed mechanism for formation of the quinone form of **1** are unknown. Under the conditions of our experiment, a $\Delta\psi$ is produced in the presence of Ca^{2+} , and not in its absence. This demonstrates that there is no competitive transmembrane transport of other ions in either direction before reaching the limiting $\Delta\psi$. We note that **1** is not observed to function as a proton pump in the presence or absence of Ca^{2+} . In contrast to the earlier-reported proton pump¹⁰, shuttle **1** was introduced into the system only in the hydroquinone form, which does not bind additional protons. The facile reduction of the quinone form of **1**, the pKa values of its various redox states, and their membrane solubility properties may play a role in the lack of proton transport.

These results demonstrate that a light-powered artificial photosynthetic reaction centre can be coupled to endergonic transmembrane Ca^{2+} transport by a redox loop to store energy as a combination of electrical and chemical potential. Although it is biomimetic only in the broadest sense, the system illustrates how various elements of biological energy transduction may be abstracted and combined in nanoscale synthetic constructs that perform interesting and potentially useful functions. □

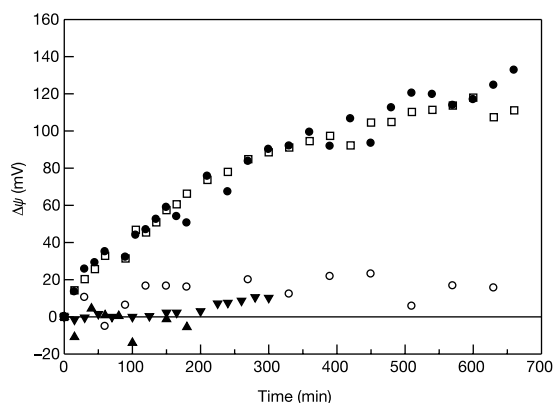


Figure 3 Light-induced formation of a membrane potential. The buildup of $\Delta\psi$ with irradiation of the liposomal system in Fig. 2 containing 1.0 mM Ca^{2+} in the external volume, and no added calcium or Fluo-3 in the internal liposomal volume, was determined from the calibrated fluorescence increase of the $\Delta\psi$ sensitive dye 8-anilino-1-naphthalenesulphonic acid (7 μM) (open squares). Also shown is the development of $\Delta\psi$ with (filled circles) and without (open circles) irradiation under the same conditions, but with 1.0 mM Ca^{2+} both inside and outside the liposomes. Control experiments include illumination by actinic light but without **2** (filled down triangles), or with replacement of **1** by its 5-methoxy analogue (filled up triangles).

Received 18 July; accepted 10 October 2002; doi:10.1038/nature01209.

1. Bartsch, R. A. & Way, J. D. *Chemical Separations with Liquid Membranes* (American Chemical Society, Washington, 1996).
2. Hirose, T., Baldwin, B. W. & Wang, Z. H. Kemp's triacid derivatives as versatile divalent metal ion transport agents and their utilization as chromoionophores. *Recent Res. Dev. Pure Appl. Chem.* **2**, 527–546 (1998).
3. de Gyves, J. & de Miguel, E. R. Metal ion separations by supported liquid membranes. *Indust. Eng. Chem. Res.* **38**, 2182–2202 (1999).
4. Lamrabet, A., Janot, J. M., Bienvenue, E., Momenteau, M. & Seta, P. Photoinitiated vectorial transmembrane electron transfer in bilayers sensitized by a face to face triporphyrin acting as a molecular electronic device. Amplification due to ionic coupling. *Photochem. Photobiol.* **54**, 123–126 (1991).
5. Zhao, Z.-G. & Tollin, G. Chlorophyll photosensitized electron transfer reactions in lipid bilayer vesicles: Generation of proton gradients across the bilayer coupled to quinone reduction and hydroquinone oxidation. *Photochem. Photobiol.* **55**, 611–619 (1992).
6. Dadfarnia, S., Shamsipur, M., Tamaddon, F. & Sharghi, H. Extraction and membrane transport of metal ions by some synthetic 9,10-anthraquinone and 9-anthrone derivatives. A selective system for calcium transport. *J. Membr. Sci.* **78**, 115–122 (1993).

7. Longin, T. L., Goyette, M. L. & Koval, C. A. Liquid membranes with light switches. *Chem. Innov.* **31**, 23–30 (2001).
8. Mitchell, P. Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. *Biol. Rev. Camb. Phil. Soc.* **41**, 445–502 (1966).
9. Sullivan, B. W. & Faulkner, D. J. *New Perspectives in Sponge Biology* (ed. Rützler, K.) 45–50 (Smithsonian Institution Press, Washington, 1990).
10. Steinberg-Yfrach, G. *et al.* Artificial photosynthetic reaction centers in liposomes: Photochemical generation of transmembrane proton potential. *Nature* **385**, 239–241 (1997).
11. Steinberg-Yfrach, G. *et al.* Light-driven production of ATP catalysed by F_0F_1 -ATP synthase in an artificial photosynthetic membrane. *Nature* **392**, 479–482 (1998).
12. Minta, A., Kao, J. P. Y. & Tsien, R. Y. Fluorescent indicators for cytosolic calcium based on rhodamine and fluorescein chromophores. *J. Biol. Chem.* **264**, 8171–8178 (1989).

Acknowledgements This research was supported by the US Department of Energy and the Harrington Arthritis Research Center.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.G. (e-mail: gust@asu.edu).

Climate change in the North Pacific region over the past three centuries

G. W. K. Moore*, Gerald Holdsworth† & Keith Alverson‡

* Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

† Arctic Institute of North America, University of Calgary, Calgary, Alberta T2N 1N4, Canada

‡ PAGES International Project Office, Bern CH-3011, Switzerland

The relatively short length of most instrumental climate records restricts the study of climate variability^{1,2}, and it is therefore essential to extend the record into the past with the help of proxy data. Only since the late 1940s have atmospheric data been available³ that are sufficient in quality and spatial resolution to identify the dominant patterns of climate variability, such as the Pacific North America pattern^{4,5} and the Pacific Decadal Oscillation⁶. Here we present a 301-year snow accumulation record from an ice core at a height of 5,340 m above sea level—from Mount Logan, in northwestern North America. This record shows features that are closely linked with the Pacific North America pattern for the period of instrumental data availability. Our record extends back in time to cover the period from the closing stages of the Little Ice Age to the warmest decade in the past millennium⁷. We find a positive, accelerating trend in snow accumulation after the middle of the nineteenth century. This trend is paralleled by a warming over northwestern North America which has been associated with secular changes in both the Pacific North America pattern and the Pacific Decadal Oscillation.

Mount Logan, located in the heavily glaciated Saint Elias mountains of the Yukon, is the highest mountain in Canada and is situated in a region of climatological importance. It is located at the end of the major North Pacific storm track⁸ along the main atmospheric pathway by which water vapour enters northwestern Canada⁹. In 1980, a 103-m ice core was retrieved from a high elevation site (5,340 m above sea level) on the mountain. Recent shallow coring and snow-pit sampling at the site have extended the ice-core record to the year 2000. Stratigraphic techniques^{10–12} allow for the retrieval of annual snow accumulation data extending from 1693 to 2000. The section before 1736 did not have annual snow accumulations assigned throughout owing to a lack of reliable seasonal indicators¹⁰. It has been shown that the annual snow

accumulation time series from this ice core contains information on circulation and temperature patterns throughout the North Pacific region^{13,14}.

Figure 1 shows the updated snow accumulation time series from the Mount Logan site over the period 1700–2000. The time series exhibits variability on both the interannual as well as interdecadal timescales¹³. Furthermore, Fig. 1 indicates that snow accumulation at the site has increased over this period, with much of the increase occurring during the latter half of the record. To assess the statistical significance of this increase, we tested the validity of the null hypothesis that the trend in annual snow accumulation at the site, obtained with the least-squares method, is indistinguishable from zero. To counteract the known tendency of geophysical time series to exhibit elevated power at low frequencies¹⁵, we assumed that the regression residuals are temporally correlated. As a result, we used a reduced effective sample size, which is a function of the lag-1 (year) autocorrelation of the regression residuals, in the test of statistical significance¹⁶. This test showed that from 1736–1850 the trend was not significantly different from zero. The lack of any trend during this period is also consistent with the data from 1700–35. Over the more recent portion of the record, from 1851 onwards, the trend is significant at the 95% confidence level. Furthermore, there is evidence for a recent acceleration in the rate of snow accumulation. In Table 1, we show that over the period 1948–2000, the trend was approximately four times as large as over the period 1851–2000. An even more rapid increase has occurred over the period 1976–2000.

Given the results presented in Fig. 1 and Table 1, we seek to identify trends in the region that have contributed to the observed increase in snow accumulation at the site. Figure 2 shows the trend in winter mean surface temperature in the North Pacific region as contained in the HADCRUTv data set¹⁷. Results are shown for the entire period of this data set, 1870–1999, as well as for the period 1948–99. Over the entire period, the dominant feature in this field is a warming in northwestern North America and the adjoining northeast Pacific Ocean. In addition to this feature, over the past 52 years there has also been a cooling over the North Pacific. There is also a marked increase in the rate of warming over North America in recent years^{17,18}.

Figure 3 shows the regression of the winter mean 500-mbar geopotential height, vertically integrated moisture transport and 1,000–500-mbar thickness fields from the National Center for Environmental Prediction (NCEP) reanalysis^{3,19} for the period 1948–2000 with the annual snow accumulation time series from the Mount Logan ice core. The 500-mbar pressure surface is located at the approximate elevation of the ice core site and its height is

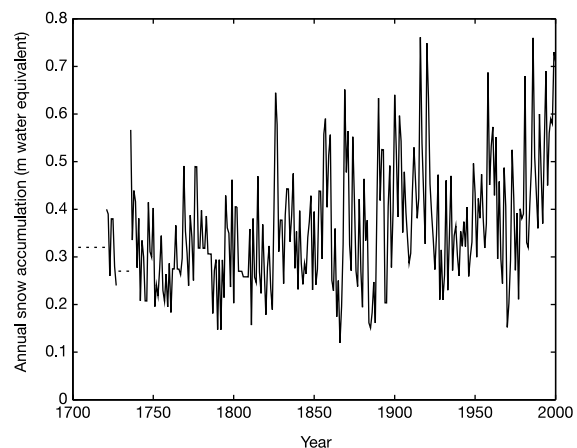


Figure 1 Annual snow accumulation (m water equivalent) at the Mount Logan site 1700–2000. Data in the pre-1736 era that is not annually resolved is indicated by the dashed lines.

widely used to provide information on the circulation in the troposphere^{4,5}. The vertically integrated moisture flux provides succinct information on the atmospheric water cycle²⁰. The thickness of the atmospheric layer between 1,000 mbar and 500 mbar is proportional to the mean temperature of the layer²¹. The main feature in the two scalar regressions is a dipolar structure with a positive pole over northwestern North America and a negative pole over the North Pacific. In particular, heavy snow accumulation at the site is associated with warmer tropospheric temperatures and higher geopotential heights over northwestern North America, with the opposite occurring over the North Pacific. Associated with these anomalies is an intensification of the eastward atmospheric moisture transport over the subtropical North Pacific and northward transport along the west coast of North America. A similar dipolar structure is also observed in the surface temperature regression (not shown) from the NCEP reanalysis over the period 1948–2000.

The dipolar structure that we have identified in our trend and regression analysis over the period 1948–2000 bears a close resemblance to the Pacific North America pattern (PNA) and the Pacific Decadal Oscillation (PDO). No evidence of this dipole is present in the surface temperature trend over the period 1870–1999 (Fig. 2a). The Pearson correlation coefficient between the annual snow accumulation time series at the Mount Logan site and winter mean indices of both the PNA⁴ and PDO⁶ provides additional evidence for this relationship. For the period 1948–2000, the correlation between the Mount Logan time series and the winter mean PNA index was 0.28, which is statistically significant at the

95% level. If we filter out high-frequency variability by using a low-pass filter with a cut-off of ten years, the correlation increases to 0.49. This correlation is statistically significant at the 99% level even when we take into account the reduction in the degrees of freedom that arises due to the low-pass filtering of the time series. It has been shown that the area-weighted sea-level pressure over the North Pacific is correlated with the PNA²². Over the period 1925–2000 (the period in which confidence in this data is highest²²), the correlation between this time series and the Mount Logan time series attains a value of 0.22 which is statistically significant at the 95% level. Over the period 1948–2000, the behaviour of the correlation between the Mount Logan time series and the winter mean PDO index was similar to that for the PNA. Over the entire period for which the PDO index is available, 1900–2000, the correlation between the PDO index and the Mount Logan time series is not statistically significant. This result is consistent with the trend data shown in Fig. 2a. Recently tree-ring data from southwestern North America have been used to develop a proxy for the PDO index²³. For the

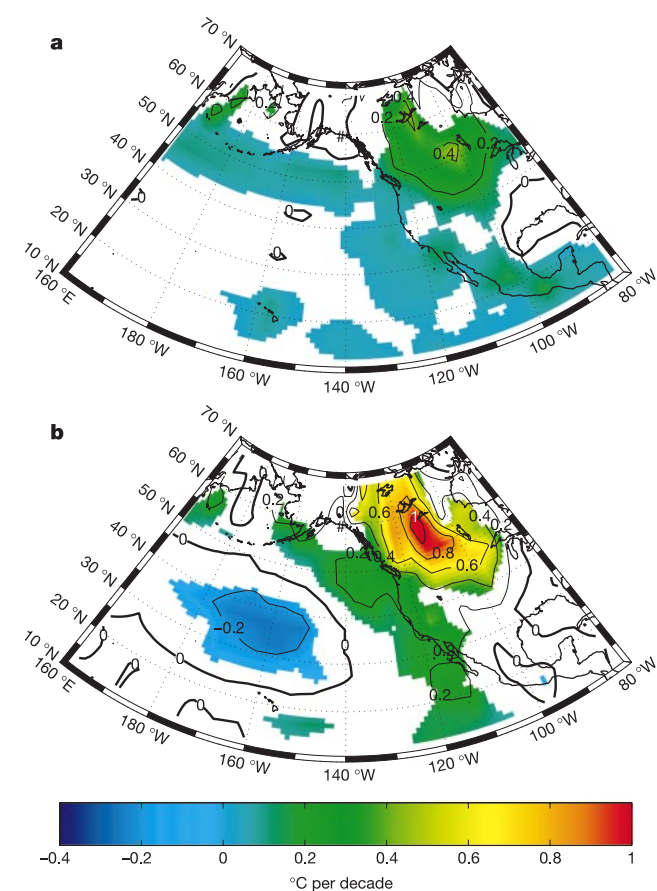


Figure 2 Trend in the winter mean (January, February, March) surface temperature field from the HADCRUTv data set. **a**, Trend during the period 1870–1999. **b**, Trend during the period 1948–1999. In each instance, shading indicates where the trend is significant at the 95% significance level in the presence of temporally autocorrelated noise. The hash symbol indicates the location of Mount Logan.

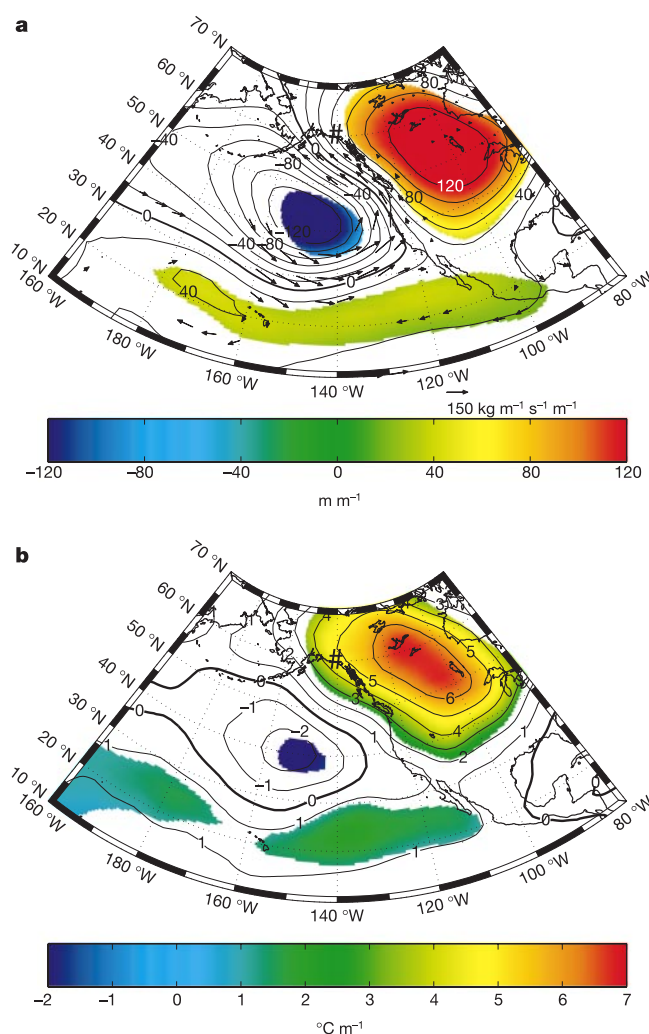


Figure 3 Regression of winter mean (January, February, March) fields from the NCEP re-analysis against the Mount Logan annual snow-accumulation time series 1948–2000. **a**, Regression of the 500-mbar geopotential height field (m m^{-1}) and the vertically integrated moisture transport vector field ($\text{kg m}^{-1} \text{s}^{-1} \text{m}^{-1}$). **b**, Regression of the mean temperature in the 1,000–500-mbar layer ($^{\circ}\text{C m}^{-1}$). For the scalar fields, shading indicates where the regression is significant at the 95% significance level. The vector field is shown at those gridpoints where at least one of its components is significant at the 95% level. The hash symbol indicates the location of Mount Logan.

Table 1 Trends in annual snow accumulation at the Mount Logan ice core site

Time interval	1736–1850	1851–2000	1948–2000	1976–2000
Trend (m per decade)	0.003	0.008	0.034	0.12
Significance level (%)	50	95	95	99

period of overlap between it and the Mount Logan time series, 1736–1991, there is no statistically significant correlation between them.

The complex relationship between the Mount Logan time series, the PNA and the PDO identified in this paper is not unique. For example, the correlation between salmon productivity in Alaska and the PDO²⁴, although robust over the past 50 years, shows changes in correlation and phase when 300-year proxy reconstructions of North Pacific sea surface temperature and salmon abundance are compared²⁵. This complexity serves to highlight the need to constrain better the temporal and spatial variability of climatic modes, such as the PDO and PNA, using several independent proxies in order to better predict their evolution and their relevance to societal concerns^{2,26}.

The three-century-long snow-accumulation record from the Mount Logan site extends from the closing stages of the Little Ice Age to the warmest decade in the past millennium⁷. A statistically significant and accelerating positive trend has existed in this time series since the middle of the nineteenth century. This was preceded by a period of approximately 150 years in which there was no trend. The results that we have presented suggest that this trend is associated with a surface warming over western North America throughout the period for which we have good-quality surface data, 1870–2000. Furthermore, the Mount Logan time series is correlated with the PNA over the period for which we have good-quality information on its variability, 1925–2000. With respect to the PDO, there exists a statistically significant correlation only after the middle of the twentieth century. It therefore appears that the secular trend in the Mount Logan snow accumulation time series is associated with a long-term intensification of the PNA. Our analysis suggests that the PDO has undergone a similar intensification, but only since the middle of the twentieth century. Over the past 50 years, this relationship between the PNA and the PDO may have contributed to the observed rapid increase in snow accumulation at the Mount Logan site and the acceleration in the warming over northwestern North America. □

Received 3 April; accepted 22 October 2002; doi:10.1038/nature01229.

1. Wunsch, C. The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Meteorol. Soc.* **80**, 245–255 (1999).
2. Alverson, K. *et al.* A global paleoclimate observing system. *Science* **293**, 47–48 (2001).
3. Kistler, R. *et al.* The NCEP-NCAR50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Am. Meteorol. Soc.* **82**, 247–267 (2001).
4. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weath. Rev.* **109**, 784–812 (1981).
5. Barnston, A. G. & Livezey, R. E. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Weath. Rev.* **115**, 1083–1126 (1987).
6. Zhang, Y., Wallace, J. M. & Battisti, D. S. ENSO-like interdecadal variability: 1900–93. *J. Clim.* **10**, 1004–1020 (1997).
7. Esper, J., Cook, E. R. & Schweingruber, F. H. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295**, 2250–2253 (2002).
8. Blackmon, M. L. Climatological spectral study of 500 Mb geopotential height of northern hemisphere. *J. Atmos. Sci.* **33**, 1607–1623 (1976).
9. Smirnov, V. V. & Moore, G. W. K. Spatial and temporal structure of atmospheric water vapor transport in the Mackenzie River basin. *J. Clim.* **12**, 681–696 (1999).
10. Holdsworth, G., Krouse, H. R. & Nosal, M. in *Climate since A.D. 1500* (eds Bradley, R. S. & Jones, P. D.) (Routledge, 1992).
11. Whitlow, S., Mayewski, P., Dibb, J., Holdsworth, G. & Twickler, M. An ice-core-based record of biomass burning in the Arctic and Sub-Arctic, 1750–1980. *Tellus Ser. B* **46**, 234–242 (1994).
12. Mayewski, P. A. *et al.* Ice-core sulfate from 3 northern-hemisphere sites—Source and temperature forcing implications. *Atmos. Environ. Part A* **27**, 2915–2919 (1993).
13. Moore, G. W. K., Holdsworth, G. & Alverson, K. Extra-tropical response to ENSO as expressed in an ice core from the Saint Elias Mountain range. *Geophys. Res. Lett.* **28**, 3457–3460 (2001).
14. Moore, G. W. K., Alverson, K. & Holdsworth, G. Variability in the climate of the Pacific Ocean and North America as expressed in an ice core from Mount Logan. *Ann. Glaciol.* (2002) (in the press).
15. Mann, M. E. & Lees, J. Robust estimation of background noise and signal detection in climatic time-series. *Clim. Change* **33**, 409–445 (1996).

16. Santer, B. D. *et al.* Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *J. Geophys. Res. Atmos.* **105**, 7337–7356 (2000).
17. Jones, P. D., New, M., Parker, D. E., Martin, S. & Rigor, I. G. Surface air temperature and its changes over the past 150 years. *Rev. Geophys.* **37**, 173–199 (1999).
18. Houghton, J. T. *et al.* (eds) *IPPC Third Assessment Report, Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, 2001).
19. Kalnay, E. *et al.* The NCEP/NCAR40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
20. Peixoto, J. P. & Oort, A. H. *Physics of Climate* (American Institute of Physics, 1992).
21. Wallace, J. M., Zhang, Y. & Renwick, J. A. Dynamic contribution to hemispheric mean temperature trends. *Science* **270**, 780–783 (1995).
22. Trenberth, K. E. Recent observed interdecadal climate changes in the northern hemisphere. *Bull. Am. Meteorol. Soc.* **71**, 988–993 (1990).
23. Biondi, F., Gershunov, A. & Cayan, D. R. North Pacific decadal climate variability since 1661. *J. Clim.* **14**, 5–10 (2001).
24. Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M. & Francis, R. C. A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Am. Meteorol. Soc.* **78**, 1069–1079 (1997).
25. Finney, B. P., Gregory-Eaves, I., Sweetman, J., Douglas, M. S. V. & Smol, J. P. Impacts of climatic change and fishing on Pacific salmon abundance over the past 300 years. *Science* **290**, 795–799 (2000).
26. Mann, M. E. The value of multiple proxies. *Science* **297**, 1481–1482 (2002).

Acknowledgements The Mount Logan time series update was supported by the International Arctic Research Center, Fairbanks, the Geological Survey of Canada, the National Institute of Polar Research, Tokyo, and the Natural Sciences and Engineering Research Council of Canada. G.H. thanks M.N. Demuth, K. Supeene, E.J. Steig and S. Rupper for assistance in acquiring and analysing the ice core. The NCEP reanalysis data was provided by the Climate Diagnostics Center of the US National Oceanic and Atmospheric Administration. The HADCRUTv data was provided by the Climate Research Unit at the University of East Anglia.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.W.K.M. (e-mail: moore@atmosph.physics.utoronto.ca).

Altered performance of forest pests under atmospheres enriched by CO₂ and O₃

Kevin E. Percy*, Caroline S. Awmack†, Richard L. Lindroth†, Mark E. Kubiske‡, Brian J. Kopper†, J. G. Isebrands§, Kurt S. Pregitzer||, George R. Hendrey¶, Richard E. Dickson‡, Donald R. Zak#, Elina Oksanen☆, Jaak Sober||, Richard Harrington** & David F. Karnosky||

* Natural Resources Canada, Canadian Forest Service-Atlantic Forestry Centre, Fredericton, New Brunswick, E3B 5P7, Canada

† Department of Entomology, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA

‡ North Central Research Station, US Forest Service, Rhinelander, Wisconsin 54501, USA

§ E7323 Highway 54, P.O. Box 54, New London, Wisconsin 54961, USA

|| School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, Michigan 49931, USA

¶ Brookhaven National Laboratory, 1 South Technology Street, Upton, New York 11973, USA

School of Natural Resources & Environment, The University of Michigan, 430 E. University, Ann Arbor, Michigan 48109, USA

☆ Department of Ecology and Environmental Science, University of Kuopio, 70211 Kuopio, Finland

** Plant and Invertebrate Ecology Division, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK

Human activity causes increasing background concentrations of the greenhouse gases CO₂ and O₃¹. Increased levels of CO₂ can be found in all terrestrial ecosystems². Damaging O₃ concentrations currently occur over 29% of the world's temperate and subpolar forests but are predicted to affect fully 60% by 2100 (ref. 3). Although individual effects of CO₂ and O₃ on vegetation have

been widely investigated, very little is known about their interaction, and long-term studies on mature trees and higher trophic levels are extremely rare⁴. Here we present evidence from the most widely distributed North American tree species⁵, *Populus tremuloides*, showing that CO₂ and O₃, singly and in combination, affected productivity, physical and chemical leaf defences and, because of changes in plant quality, insect and disease populations. Our data show that feedbacks to plant growth from changes induced by CO₂ and O₃ in plant quality and pest performance are likely. Assessments of global change effects on forest ecosystems must therefore consider the interacting effects of CO₂ and O₃ on plant performance, as well as the implications of increased pest activity.

Atmospheric CO₂ concentrations have risen by nearly 30% since the mid-1800s (ref. 6), and background levels of tropospheric O₃ have risen concurrently from 10–15 µl l⁻¹ a century ago to 30–40 µl l⁻¹ today⁷. Controlled exposure of trees to CO₂ or O₃ drives plant responses in different directions, producing stimulatory (CO₂) or inhibitory (O₃) effects on aboveground growth⁸. Increases or decreases in substrate availability due to CO₂ and O₃ change metabolic use of growth-limiting resources such as light, water and nitrogen, and also alter the physical and chemical quality of plant tissues. In turn, such changes can affect the performance of higher trophic levels, particularly insects⁹ and plant diseases¹⁰. Field-based investigations of the impacts of CO₂- and O₃-induced changes in tree quality on natural populations of pests and diseases have, however, only become possible since the development of free-air

carbon dioxide enrichment (FACE) technology¹¹.

Here we report on the first free-air investigations into the impacts of increased CO₂ and O₃ on natural populations of forest insects and their natural enemies, and on natural outbreaks of tree diseases. The data we present link changes in physical and chemical plant defences with pest performance within the context of simultaneous treatment-induced changes to tree growth. To do this, we chose three model systems: one plant pathogen and two insects having widely different feeding strategies. The poplar leaf rust, *Melampsora medusae*, is common on aspen and belongs to the most widely occurring group of foliage diseases¹². The forest tent caterpillar, *Malacosoma disstria*, is a common leaf-chewing lepidopteran in North American hardwood forests¹³. The sap-feeding aphid, *Chaitophorus stewartii*, infests aspen throughout its range¹⁴.

Our data show that enhanced CO₂ and O₃ had no effect on trembling aspen tree height in 1998 or 1999, but that by 2000, with the onset of canopy closure, height growth began to diverge among

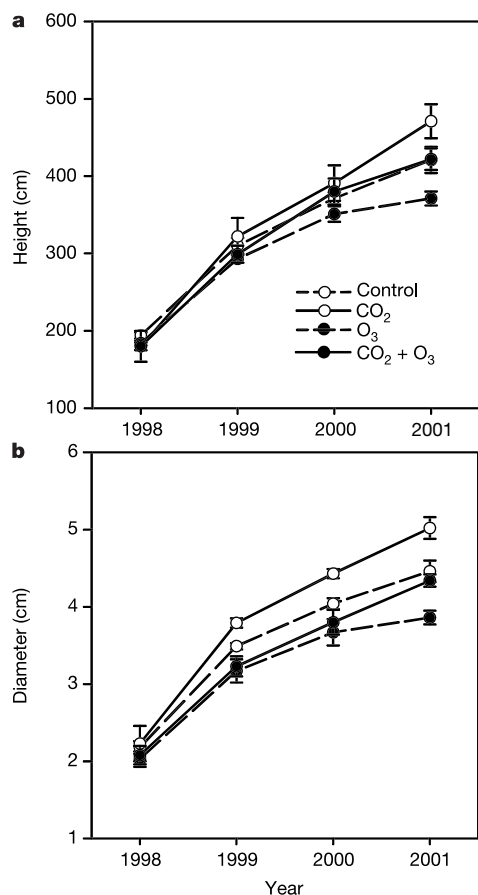


Figure 1 The effects of elevated CO₂ and O₃ on the growth of trembling aspen. Data are presented as means ± 1 s.e. **a**, We note the divergence of height due to treatment in year four of fumigation, the height stimulation due to CO₂ and the height depression due to O₃. **b**, Diameter growth due to treatment diverged earlier, after two years of fumigation.

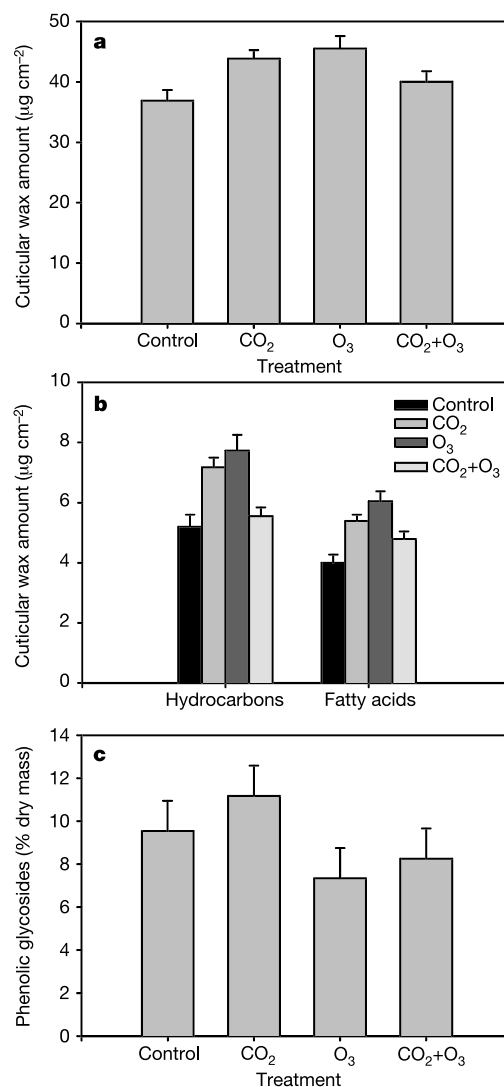


Figure 2 The effects of CO₂ and O₃ on physical and chemical leaf defences in trembling aspen. Data are presented as means ± 1 s.e. **a**, Comparison of amount of the physical barrier (cuticular wax) produced showing the contradictory influence of O₃ alone or in combination with CO₂. **b**, Comparison of carbon allocation among major cuticular wax classes known to act as phytophagous insect deterrents or stimulants. **c**, Concentrations of the important aspen defensive metabolites, phenolic glycosides (salicortin + tremulacin).

treatments (Fig. 1a). In 2001, four years after planting, trees grown in enhanced CO₂ were taller ($P = 0.005$), and those in enhanced O₃ were smaller ($P = 0.001$) than controls. Under CO₂ + O₃, the beneficial effects of CO₂ negated the adverse effects of O₃ ($P = 0.278$; Fig. 1a). Tree diameters, however, diverged almost immediately, and significant treatment effects continued through the 2001 growing season. Enhanced CO₂ increased tree diameter ($P < 0.06$), enhanced O₃ decreased it ($P < 0.06$) and trees with enhanced CO₂ + O₃ did not differ from the controls (Fig. 1b). These four-year data support our premise that, based upon a 2100 scenario for CO₂ and O₃, aspen growth, survival and productivity will be greatly affected by these two air pollutants¹⁵. The implications are huge because, since the mid-1990s, aspen use has increased almost exponentially in the US Great Lakes Region and in the boreal mixed wood region of Canada, where the resource is now almost fully utilized¹⁶.

Because our tree performance data were gathered from a 'natural'

ecosystem, we were able to monitor natural outbreaks of pests and diseases. We relate the impacts of CO₂ and O₃ on populations of insects and diseases to those on trees in the control rings (36011⁻¹ CO₂; 36.0–38.8 μl l⁻¹ O₃ during the growing season (1998–2001) in the daytime (07:00–18:59) on average) because there are no 'pest-free' trees in our experiment.

We focused our initial effort on the leaf cuticle, an important defensive barrier against pests¹⁷. Exposure of aspen to enhanced CO₂ and/or O₃ greatly altered production ($P < 0.0001$) and chemistry

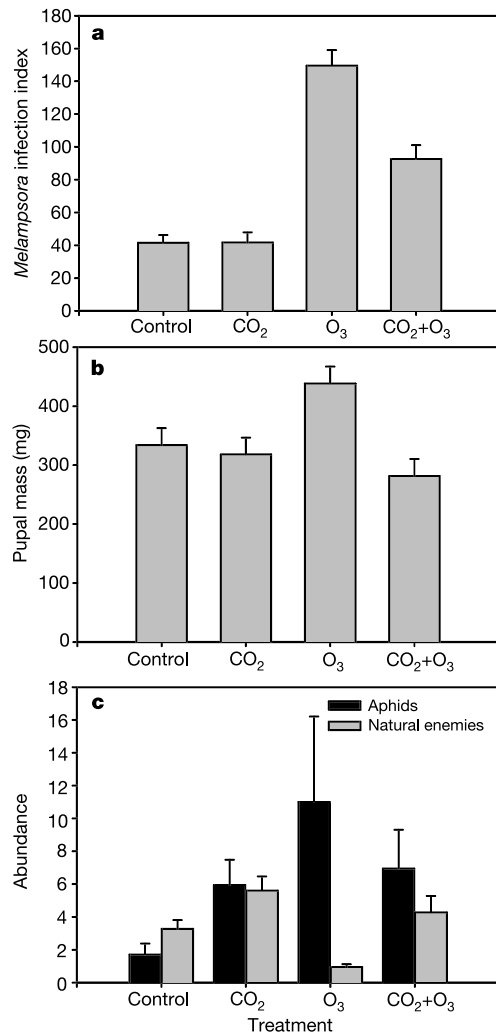


Figure 3 Effects of CO₂ and O₃ on one pathogen and two important herbivorous insects found on aspen **a**, *Melampsora medusae* infection; we note that rust increased on leaves exposed to O₃ relative to both the control and those exposed to CO₂ alone. A mid-range response was observed for leaves exposed to CO₂ + O₃. **b**, Performance of a lepidopteran, leaf-chewing insect, the forest tent caterpillar, as assessed by female pupal mass. We note the increased performance under O₃ and decreased performance under CO₂, especially under high O₃. **c**, Mean numbers of aphids and natural enemies (ladybirds, lacewings, spiders and parasitoids) per tree. Under O₃, sap-feeding aphid numbers increased while their natural enemies simultaneously decreased.

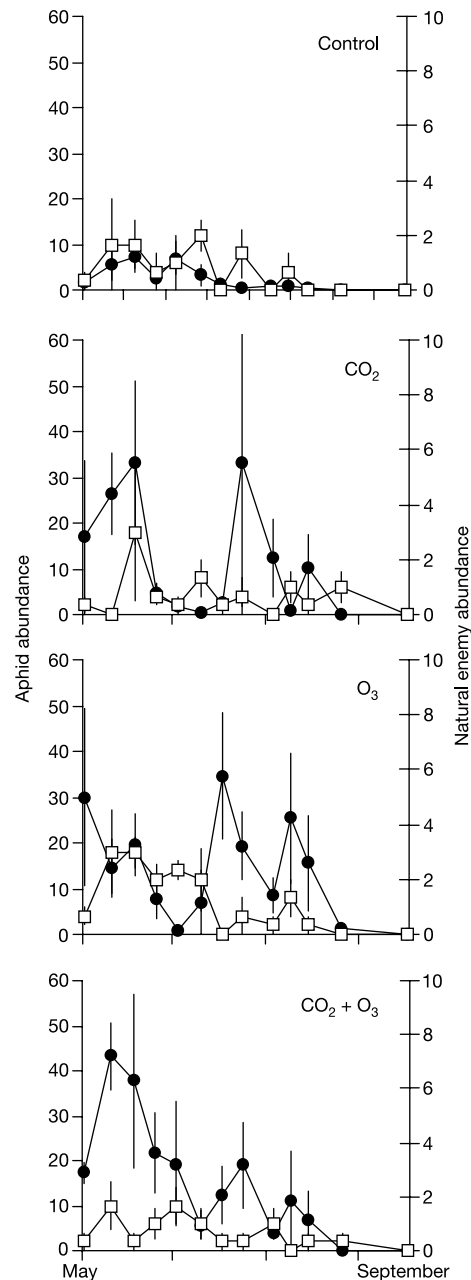


Figure 4 Long-term effects of enhanced CO₂ and O₃ atmospheres on the abundance of aphids (filled circles) and natural enemies (open squares) feeding on aspen. Aphid abundance increased significantly under elevated CO₂ and/or O₃, while natural enemy abundance did not, although we note changes in the patterns of natural enemy population increases.

($P < 0.001$) of cuticular waxes. Exposure to CO_2 or O_3 stimulated cuticular-wax production (16 and 23%, respectively) (Fig. 2a). Increasing C allocation to cuticular waxes owing to CO_2 ($P = 0.0029$) or O_3 ($P = 0.0021$), however, was negated under $\text{CO}_2 + \text{O}_3$ ($P = 0.215$) (Fig. 2a). Of particular relevance is that modifications to cuticular waxes determine the physicochemical characteristics of the plant surface on a scale relevant to plant-eating and insect-eating insects through action upon attachment, host location and host acceptance¹⁸. We show here that synthesis of fatty acids (which stimulate host recognition¹⁸) was increased by CO_2 ($P = 0.0001$) and O_3 ($P = 0.0001$) applied singly, as well as by $\text{CO}_2 + \text{O}_3$ ($P = 0.037$) (Fig. 2b). Coincidentally, both CO_2 ($P = 0.0002$) and O_3 ($P = 0.0002$) increased amounts of hydrocarbons (Fig. 2b), which are known to help insects recognize surfaces by increasing their responsiveness to chemical cues¹⁸.

Changes to plant defences under enhanced CO_2 and O_3 also involve changes in the chemical composition of the leaf tissue. Trembling aspen accumulates high concentrations of phenolic glycosides, which are of singular importance as protective agents against pests¹⁹. Phenolic glycoside concentrations increased under enhanced CO_2 ($P = 0.046$), and decreased under enhanced O_3 ($P = 0.002$) (Fig. 2c). The independent effects of CO_2 and O_3 were attenuated when the trees were exposed to both gases simultaneously. As CO_2 and O_3 obviously induced changes to aspen physical and chemical defensive characteristics, we next investigated the impacts of these changes in host quality on performance of the insects and diseases naturally infesting our aspen trees.

The role of cuticular wax as a physical barrier to protect leaves from fungal infection is well established²⁰. We have previously shown how O_3 alters aspen cuticular-wax structure and increases leaf wetting²¹; the latter is a function of the ratio of the most hydrophobic (hydrocarbons) to least (fatty acids) hydrophobic cuticular wax²². Rust infection increased fourfold under enhanced O_3 ($P < 0.001$, Fig. 3a). CO_2 alone did not alter rust occurrence. Interestingly, co-exposure did not completely offset the negative effects of O_3 , as infection remained almost threefold increased compared with the control leaves ($P = 0.022$). As yield loss due to *Melampsora medusae* can lead to reduction in diameter growth and delayed development of root systems²³, we suggest that the negative effects of O_3 environments on aspen productivity in Aspen-FACE (Fig. 1) were due to the interacting stresses of O_3 toxicity and increased leaf damage by *M. medusae*.

Historically, the forest tent caterpillar has defoliated more deciduous forest than any other insect in North America and regularly defoliates large areas (up to 26×10^6 ha) of early successional forests¹⁶. Outbreaks can reduce timber yield up to 90% in one year, and increase tree vulnerability to disease and environmental stress²⁴. Enhanced CO_2 ($P = 0.020$) reduced female pupal mass, particularly under high levels of O_3 ($\text{CO}_2 \times \text{O}_3$ interaction, $P = 0.046$) (Fig. 3b). In contrast, O_3 fumigation improved tent caterpillar performance, increasing female pupal mass by 31% ($P = 0.039$). As large forest tent caterpillars produce more offspring than small individuals²⁵, the increase in pupal mass due to O_3 suggests that future outbreaks of this important forest pest will be more severe (under higher O_3 environments) than at present.

Changes in forest tent caterpillar performance may also affect interactions with higher trophic levels, because natural enemy performance frequently depends on size and quality of the prey⁹. We chose aphids (and their natural enemies) for our population surveys of tritrophic interactions. Aphids were ideal for investigations of the impacts of CO_2 and O_3 on pest and natural enemy populations because aphids spend their entire life cycle on the trees in the FACE rings and the performance of multiple generations can be investigated during a single growing season. Population surveys revealed that, although aphid abundance was unaffected by CO_2 and O_3 (Fig. 3c), natural enemy densities were much larger ($P < 0.001$) under CO_2 than under the control treatment. In

contrast, O_3 had a strong negative effect ($P = 0.002$) on natural enemies, but did not modify CO_2 effects ($P = 0.241$).

The long-term population surveys we conducted the following year showed that, despite the high variability associated with natural aphid infestations (a result of their exponential population growth rates), aphid abundance was, on average, significantly ($P = 0.0231$) greater on trees grown under enhanced O_3 than on control trees (Fig. 4). Peak aphid populations also increased on trees grown under enhanced CO_2 ; although in this case, the result was dependent on the sampling date and O_3 concentration (significant $\text{CO}_2 \times \text{O}_3 \times \text{date}$ interaction; $P = 0.019$). In this long-term study, enhanced CO_2 and O_3 had no overall effect on the abundance of natural enemies (averaged over the growing season), but did affect the synchrony of natural enemies with their aphid hosts (Fig. 4). The long-term investigation of the impacts of enhanced CO_2 and O_3 on the population dynamics of aphids and their natural enemies clearly shows that aphid infestations were more severe on trees grown under enhanced CO_2 , O_3 or $\text{CO}_2 + \text{O}_3$ atmospheres. Aphid populations under enhanced CO_2 and O_3 did not reflect the performance of individual aphids (which was unaffected by CO_2 and O_3) or the effects of CO_2 and O_3 on the nutritional or defensive quality of the leaf tissues (Fig. 2c); as aphids are phloem-feeders, this was not surprising. The effects of CO_2 and O_3 on aphid abundance did, however, correlate well with the increase in plant cuticular wax (Fig. 2a, b) and also reflected trends observed for *M. medusae* (Fig. 3a).

Populus tremuloides and its close relative European aspen, *Populus tremula*, are worldwide species, and important components of many forest ecosystems⁵. Our report represents the first ecosystem-level description of the combined effects of enhanced CO_2 and O_3 on the relationship between food quality and the long-term population dynamics of major forest pests. Here, we have shown the potential for a multitrophic effect, albeit for selected plant defences and pest taxa, within an ecologically important forest ecosystem grown under realistic exposure to future CO_2 and O_3 concentrations. From our results, we believe it important to emphasize that: (1) studies of the performance of individual herbivorous insects would not have allowed us to predict the long-term population dynamics of pest and natural enemy populations; (2) although natural enemy abundance may be affected by changes in CO_2 and O_3 , the long-term regulation of pest populations by natural enemies appears to be less effective under enhanced CO_2 and/or O_3 ; (3) changes in pest dynamics reflect changes in leaf defensive qualities, particularly aspects of leaf quality such as changes in 'host signalling' at the leaf surface, with wider implications for host 'acceptance' by other pest/pathogen species; and (4) feedbacks to tree productivity are likely now occurring or will ensue and may modify forest responses to increasing CO_2 and O_3 .

In its Third Assessment Report, the IPCC stated that "...the increase of greenhouse gas concentrations in the atmosphere can be reduced also by enhanced uptake of carbon through, for example, afforestation, reforestation, slowing deforestation, and improved forest, rangeland, wetland, and cropland management..."²⁶. Yet the degree to which terrestrial ecosystems will continue to be net sinks for carbon is highly uncertain, owing to the complex interactions among an array of natural and non-natural factors. We believe the assessments of carbon sequestration in woody biomass²⁷, without consideration of the importance of CO_2 and O_3 interactions and the potential for increases in populations of insects and diseases, will seriously overestimate the positive effects of enhanced CO_2 atmospheres on the productivity of northern forests. □

Methods

The experiment

The Aspen-FACE site (32 ha) is located on a level to gently rolling sandy loam in northern Wisconsin, near Rhinelander (longitude, 89.7°, latitude, 45.6°), well within the natural range of the dominant experimental species, trembling aspen, *Populus tremuloides*

Michx.²⁸ The FACE system was constructed in 1997, and has been used to fumigate (1998–2001) developing forest stands. The experiment consists of 12, 30-m-diameter FACE rings, assigned to factorial treatments of atmospheric CO₂ (ambient and 560 l l⁻¹ during daytime hours) and O₃ (ambient and 46.4 μl l⁻¹ to 55.5 μl l⁻¹ during the growing season (1998–2001) in the daytime (07:00–18:59) on average for the six O₃ treatment rings). The four treatments are arranged in a randomized complete block design, with three replications of each treatment. Each FACE ring is divided by a walkway system into three parts. In one half of each ring, we planted five trembling aspen genotypes of differing O₃ and CO₂ responsiveness. The other half of each ring is further divided into two quarters; one is planted with aspen and sugar maple, *Acer saccharum* Marsh., and the other is planted with aspen and paper birch, *Betula papyrifera* Marsh.; each FACE ring is planted at a 1 × 1 m spacing. In June 1997, 12,000 trees were planted inside the rings (1,000 trees per ring). The performance of our FACE exposure system, the growth responses of the trees and additional details of our experiment have been summarized in detail elsewhere²⁸.

Tree height and diameter measurement

Tree height and diameter were measured at the end of each growing season, after leaf fall. Tree height was assessed using a telescoping height pole and diameter was assessed at 3 cm above the soil surface along two axes, using callipers¹⁵.

Plant assay methods

Cuticular waxes were recovered by chloroform washing from physiologically similar, fully expanded main lateral leaves collected from aspen trees growing within each ring. Wax amount (10 ± μg) was expressed in relation to leaf surface area and quantitative leaf cuticular wax chemical composition was determined using gas chromatography (GC) and GC-mass spectrometry (MS) as described elsewhere²⁹.

Positive identification of *Melampsora medusae* Thum. F. sp. *tremuloidae* was confirmed by observations of urediniospores using scanning electron microscopy (SEM) and/or light microscopy. Rust incidence was assessed in late September as percentage of infected leaves per tree and severity of rust occurrence per leaf. Their product was calculated as the incidence of infection¹⁰.

Invertebrate surveys

Leaves used for phenolic glycosides determinations were collected during the fourth larval instar of forest tent caterpillar development. Foliage (2–3 g fresh mass) was excised at the petiole from each tree, flash frozen in liquid nitrogen, freeze-dried, ground, and stored at -20°C before analysis. Levels of the phenolic glycosides salicortin and tremulacin were measured using high-performance thin-layer chromatography (HPTLC), with purified aspen salicortin and tremulacin as reference standards³⁰. Forest tent caterpillar larvae were reared communally in large mesh bags until the fourth instar, at which time they were placed in individual mesh bags (five larvae per tree, three trees per ring) and reared to pupation.

Total numbers of aphids and natural enemies on seven aspen trees from each ring were counted. Data were log₁₀ (n + 1) transformed before analysis. No winged aphids were found on any of the trees; thus all aphids are likely to have been born and developed from the initial colonization of the trees by winged females early in the growing season. The natural enemies were mainly predators (such as Coleoptera and Neuroptera) and parasitic Hymenoptera. Although the data are likely to have underestimated the total number of natural enemies (particularly winged species) exploiting the aphid populations, they represent a snapshot of the population at the time of the survey. More comprehensive sampling (for example, by using insect traps) would have included natural enemies exploiting insects feeding on all the trees growing in the FACE rings, rather than aspen-feeding aphids. The long-term aphid and natural enemy population surveys used similar methods, except that data were collected from the lowest south-facing branch on each tree rather than from entire trees. In the previous survey, aspen aphids were found only on the lower branches; no aphids were ever found on the upper branches.

Statistical considerations

Data were analysed by general linear models analyses of variance (ANOVA). The Aspen-FACE experimental design, at the whole-plot (error d.f. = 6) level, is three replications of a randomized complete block design with four treatment combinations. We assume that replications are random effects, and treatments are fixed effects²⁸.

Received 17 April; accepted 11 July 2002; doi:10.1038/nature01028.

1. IPCC *Climate Change 2001: The Scientific Basis* (Report of Working Group I of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 2001); available at <http://www.ipcc.ch/pub/spm22-01.pdf>.
2. Ciais, P., Peylin, P. & Bousquet, P. Regional biospheric fluxes as inferred from atmospheric CO₂ measurements. *Ecol. Appl.* **10**, 1574–1589 (2000).
3. Fowler, D. *et al.* The global exposure of forests to air pollutants. *Wat. Air Soil Pollut.* **116**, 5–32 (1999).
4. McLaughlin, S. B. & Percy, K. E. Forest health in North America: Some perspectives on actual and potential roles of climate and air pollution. *Wat. Air Soil Pollut.* **116**, 151–197 (1999).
5. Perala, D. A. *Silvics of North America Hardwoods* (eds Burns, R. M. & Honkala, B. H.) Vol. 2 555–569 (USDA Forest Service, Washington, DC, 1990).
6. Barnola, J. M. *et al.* CO₂ evolution during the last millennium as recorded by Antarctic and Greenland ice. *Tellus B* **47**, 264–272 (1995).
7. Finlayson-Pitts, B. J. & Pitts, J. N. Jr *Chemistry of the Upper and Lower Atmosphere* (Academic, San Diego, CA, 1999).
8. Dickson, R. E. *et al.* Growth of five hybrid poplar genotypes exposed to interacting elevated CO₂ and O₃. *Can. J. Forest Res.* **28**, 1706–1716 (1998).
9. Awmack, C. S. & Leather, S. R. Host plant quality and fecundity in herbivorous insects. *Annu. Rev. Entomol.* **47**, 817–844 (2002).
10. Karnosky, D. F. *et al.* Interacting CO₂, tropospheric O₃ and predisposition of aspen *Populus tremuloides* Michx. to attack by *Melampsora medusae* rust. *Glob. Change Biol.* **8**, 329–338 (2002).

11. Hendrey, G. R., Ellsworth, D. S., Lewin, K. F. & Nagy, J. A free-air system for exposing tall forest vegetation to elevated atmospheric CO₂. *Glob. Change Biol.* **5**, 293–309 (1999).
12. Ziller, W. G. *The Tree Rusts of Western Canada* (Canadian Forestry Service Publication No. 1329, Victoria, British Columbia, 1974).
13. Martineau, R. *Insects Harmful to Forest Trees* (Agriculture Canada, Canadian Forest Service, Quebec, 1984).
14. USDA Forest Service *Aphids* (Forest Health Protection, Southern Region); available at <http://fhpr8.srs.fs.fed.us/identis/insects/aphids.html> (2001).
15. Isebrands, J. G. *et al.* Growth responses of *Populus tremuloides* clones to interacting carbon dioxide and tropospheric ozone. *Environ. Pollut.* **115**, 359–371 (2001).
16. David, A. J., Zasada, J. C., Gilmore, D. W. & Landäusser, S. M. Current trends in the management of aspen and mixed aspen forests for sustainable production. *Forest Chron.* **77**, 525–532 (2001).
17. Jeffree, C. E. *Plant Cuticles: An Integrated Functional Approach* (ed. Kerstiens, G.) 33–82 (Bios Scientific, Oxford, 1996).
18. Eigenbrode, S. D. *Plant Cuticles: An Integrated Functional Approach* (ed. Kerstiens, G.) 201–221 (Bios Scientific, Oxford, 1996).
19. Lindroth, R. L. & Hwang, S.-Y. Diversity, redundancy and multiplicity in chemical defense system of aspen. *Rec. Adv. Phytochem.* **30**, 25–56 (1996).
20. Mendgen, K. *Plant Cuticles: An Integrated Functional Approach* (ed. Kerstiens, G.) 175–188 (Bios Scientific, Oxford, 1996).
21. Karnosky, D. F. *et al.* Effects of tropospheric O₃ on trembling aspen and interaction with CO₂: results from an O₃-gradient and a FACE experiment. *Wat. Air Soil Pollut.* **116**, 311–322 (1999).
22. Holloway, P. J. Chemistry of leaf waxes in relation to wetting. *J. Sci. Fd Agric.* **20**, 124–128 (1969).
23. Davis, C. & Meyer, T. *Field Guide to Tree Diseases of Ontario* (Natural Resources Canada, Canadian Forest Service, Sault Ste. Marie, Ontario, 1997).
24. Fitzgerald, T. D. *The Tent Caterpillars* (Cornell Univ. Press, Ithaca, New York, 1995).
25. Parry, D., Goyer, R. A. & Lenhard, G. J. Macrogeographic clines in fecundity, reproductive allocation, and offspring size of the forest tent caterpillar *Malacosoma disstria*. *Ecol. Entomol.* **26**, 281–291 (2001).
26. IPCC *Summary for Policymakers to Climate Change 2001: Synthesis Report of the Third IPCC Assessment Report* (IPCC Secretariat, Geneva, Switzerland, 2001); available at <http://www.ipcc.ch/pub/SYR-text.pdf>.
27. Tuskan, G. A. & Walsh, M. E. Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: A U.S. case study. *Forest. Chron.* **77**, 259–264 (2001).
28. Dickson, R. E. *et al.* *Forest Atmosphere Carbon Transfer Storage-II (FACTSII)—The Aspen Free-Air CO₂ and O₃ Enrichment (FACE) Project: An Overview* (USDA Forest service General Tech. Rep. NC-214, St Paul, MN, 2000).
29. Percy, K. E., McQuattie, C. J. & Rebbeck, J. A. *Air Pollutants and the Leaf Cuticle* (eds Percy, K. E., Cape, J. N., Jagels, R. & Simpson, C. M.) 67–79 (Springer, Heidelberg, 1994).
30. Lindroth, R. L., Kinney, K. K. & Platz, C. L. Responses of deciduous trees to elevated atmospheric CO₂: productivity, phytochemistry and insect performance. *Ecology* **74**, 763–777 (1993).

Acknowledgements The FACTSII: Aspen FACE project is supported by the US Department of Energy—Office of Biological and Environmental Research (DOE-BER) through its Program for Ecosystem Research. The research described herein is supported by companion grants from the US Forest Service Northern Global Change Program, Michigan Technological University, the USDA National Research Initiatives Program, the Canadian Federal Panel on Energy Research and Development (PERD), Natural Resources Canada-Canadian Forest Service, the National Science Foundation, the Natural Environmental Research Council of the United Kingdom, the British Ecological Society and the Praxair Foundation. The Aspen FACE project is an experiment within the International Biosphere Geosphere Programme and the only open-air site in the world that combines the effects of CO₂ and O₃. It is a multi-agency, interdisciplinary project with 50 scientists from seven countries.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.E.P. (e-mail: kpercy@nrcan.gc.ca).

Genetic mechanisms of floral trait correlations in a natural population

Jeffrey K. Conner

Kellogg Biological Station and Department of Plant Biology, Michigan State University, 3700 East Gull Lake Drive, Hickory Corners, Michigan 49060, USA

Genetic correlations among traits are important in evolution, as they can constrain evolutionary change or reflect past selection for combinations of traits^{1,2}. Constraints and integration depend on whether the correlations are caused by pleiotropy or linkage disequilibrium³, but these genetic mechanisms underlying correlations remain largely unknown in natural populations⁴. Quan-

titative trait locus (QTL) mapping studies do not adequately address the mechanisms of within-population genetic correlations because they rely on crosses between distinct species, inbred lines or selected lines (see ref. 5), and they cannot distinguish moderate linkage disequilibrium from pleiotropy because they commonly rely on only one or two episodes of recombination⁶. Here I report that after nine generations of enforced random mating (nine episodes of recombination), correlations between six floral traits in wild radish plants are unchanged, showing that pleiotropy generates the correlations. There is no evidence for linkage disequilibrium despite previous correlational selection acting on one functionally integrated pair of traits⁷. This study provides direct evidence of the genetic mechanisms underlying correlations between quantitative traits in a natural population and suggests that there may be constraints on the independent evolution of pairs of highly correlated traits.

In the past two decades, biologists have become increasingly

aware of the profound effects that genetic correlations can have on evolutionary change^{3,8}. A genetic correlation between two phenotypic traits can be caused either by pleiotropy, in which one locus affects both traits, or by linkage disequilibrium, in which the two traits are affected by distinct gene loci but some evolutionary force creates and maintains a nonrandom association between the alleles present at these loci⁹. Genetic correlations are important because natural selection on one character either causes an evolutionary change in a correlated neutral character or alters the response to selection in a correlated character that is itself under direct selection. In other words, non-adaptive evolution of the second character can occur. Therefore, genetic correlations among traits can slow or constrain the evolution of the most advantageous combinations of traits, at least in the short term^{1,10}.

But constraints will only occur if the correlation is caused by pleiotropy or extremely tight linkage; correlations caused by linkage disequilibrium between moderately or loosely linked loci can be altered rapidly by recombination and selection^{4,8}. Conversely, natural selection may alter correlations, especially in cases where two or more traits interact to carry out a given function^{2,11}. Here again the genetic mechanism is important, because selection is one of the principal factors that can create and maintain linkage disequilibrium³, although selection can alter pleiotropic correlations as well. Although pleiotropy is thought to be ubiquitous and the most common mechanism of genetic correlations in nature^{11–13}, the limited information available from mapping studies suggests that both linkage disequilibrium and pleiotropy can be important^{5,14}.

I studied mechanisms of correlations among six floral traits in a natural population of wild radish, *Raphanus raphanistrum*¹⁵. Wild radish is self-incompatible and therefore an obligate outcrosser. Linkage disequilibrium will only occur in natural populations of

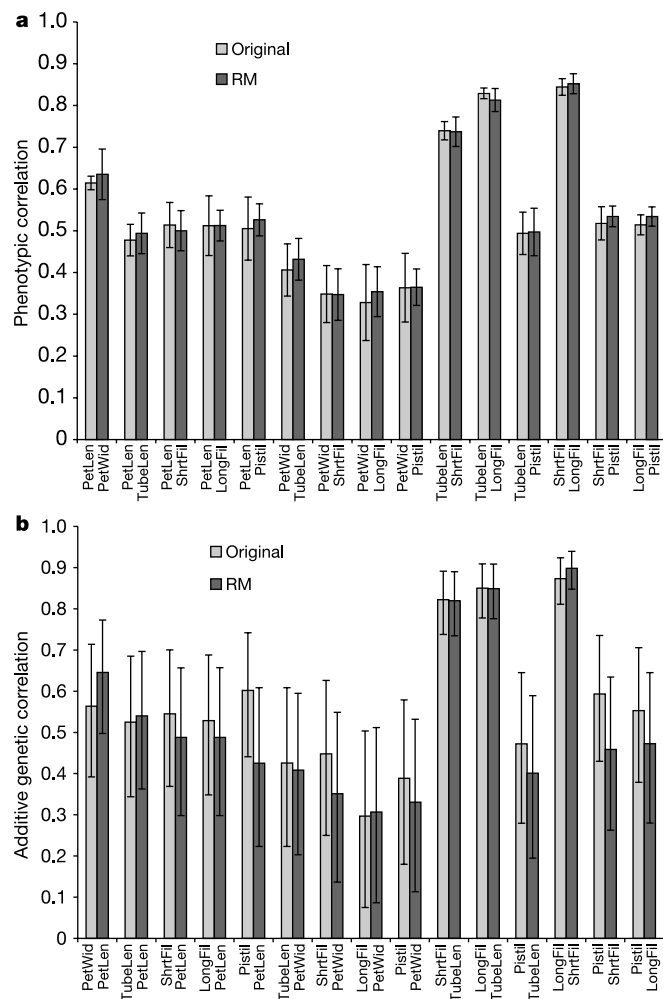


Figure 1 Floral trait correlations with 95% confidence intervals. **a**, Phenotypic correlations based on mean matrices; residual matrices gave virtually identical results (see Methods). The confidence intervals are twice the standard errors of these means. **b**, Additive genetic correlations, calculated as correlations among breeding values determined using BLUP; component correlations gave virtually identical results (see Methods). Confidence intervals are from the z transformation²⁹. The RM data are the average of the two random-mated replicates. PetLen, petal length; PetWid, petal width; TubeLen, corolla tube length; ShrtFil, short filament length; LongFil, long filament length; Pistil, pistil length.

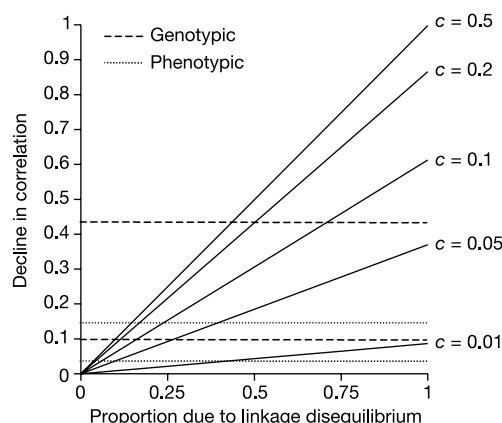


Figure 2 Theoretical declines in correlations caused by a reduction in linkage disequilibrium during nine generations of random mating. The magnitude of the reduction in a correlation is plotted on the y axis, the proportion of the original correlation that was due to linkage disequilibrium, that is, one minus the proportion owing to pleiotropy, is plotted on the x axis. Note that intermediate values for this proportion assume that several loci are affecting the correlation. The plotted lines show the declines over nine generations for five different rates of recombination from $1 - (1 - c)^9$, where c is the recombination rate³. For polygenic traits, c is a weighted average of the recombination rates between the various loci that affect the traits³⁰. Dotted and dashed lines show the range of detectable changes in phenotypic and genetic correlations, respectively. The upper dotted and dashed lines are the detectable declines for the correlation with the largest combined 95% confidence interval (CI) in the experiment (long filament and petal width, Fig. 1); that is, with a decline of 0.15 [(original CI + RM CI)/2 = 0.15] in the phenotypic correlation or 0.43 in the genetic correlation, the 95% confidence intervals would not overlap. The lower dotted and dashed lines are the detectable declines for the correlations with the smallest confidence intervals in the experiment (the correlations between the long filament, the short filament and corolla tube lengths).

outcrossers if some force such as selection or drift counteracts recombination. The phenotypic and genetic correlation between filament and corolla tube length in wild radish is much greater in magnitude than are other floral correlations in this species¹⁵, and there is much evidence for correlational selection acting on these traits^{7,16}. Therefore, linkage disequilibrium is a plausible mechanism for at least some of the filament–corolla tube correlation. The corolla and stamens (including the filaments) have been shown to have a very close developmental relationship in many plant species¹⁷, which might suggest that the higher correlation is due solely to pleiotropy. Although some other species in the Brassicaceae (the family to which wild radish belongs) share the very high filament–corolla tube correlation, many do not¹⁶, indicating that a close developmental relationship may not necessarily lead to a high correlation, or that the close developmental relationship is evolutionarily labile.

I used seeds from 600 randomly selected field plants (see ref. 15 for details) to create two greenhouse populations of 300 plants each (RM1 and RM2). In each generation, crosses were carried out randomly so that every plant acted as a male and a female, but with no reciprocal crosses. One plant from each maternal plant was raised to form the next generation. In this mating design there is no selection, because all plants have equal fitness. Under such relaxed selection, recombination will reduce the contribution of linkage disequilibrium to genetic correlations, whereas the contribution of pleiotropy will persist.

My design is similar in many respects to one used to measure mutation accumulation¹⁸. The effects of mutation accumulation should be minimal in my study because, first, floral traits were measured rather than fitness, which means that there should be far fewer genes involved; second, the traits were measured under benign conditions, which minimizes the phenotypic effects of mutations (ref. 18 and references therein); and third, the study only lasted nine generations. The absence of mutational effects is supported by the lack of difference between the mean values of the traits in the random-mated and those of the control populations (see below).

After nine generations of random mating (that is, nine episodes of recombination), 275–298 plants were raised from each of the two randomly mated populations and from a random sample of the original field population; the latter were from seeds stored from the same field collection used to establish the random mating populations. In each group, 68–75 plants were chosen randomly to be sires, and each was used to pollinate 1–3 unique randomly chosen plants (a total of 212 sires and 601 dams). One seed from each dam was raised in the greenhouse in each of two blocks separated in time for a total of 1,172 offspring (30 offspring failed to germinate). In both parental and offspring generations, the three groups were interspersed in a systematic fashion on the greenhouse bench, eliminating any average environmental differences between the groups. Six floral dimensions (petal width and lengths of petal, corolla tube, short and long filaments, and pistil) were measured on 1 of the first 40 flowers produced by each plant in both generations (93% of these were the third to thirteenth flower; see ref. 15 for details).

Owing to the design of this study and previous work on this species, differences in the phenotypic correlations among treatment groups are likely to reflect accurately differences in the additive genetic correlations. The phenotypic analyses have greater power than have genetic analyses, because power for the former depends on all individuals measured, whereas the number of families (always a smaller number) is more crucial for the latter¹⁹. But it is possible that environmental or non-additive genetic variances and covariances differ among the groups in ways that would create phenotypic differences without additive genetic differences or obscure phenotypic differences even if genetic changes had occurred. For these reasons I present both additive genetic and phenotypic analyses, which are in excellent agreement with each other—there was no

evidence for changes in the means, variances, covariances or correlations among traits owing to the nine generations of recombination.

Multivariate analysis of variance showed no evidence for differences in the mean values of the traits among the three treatment groups. Although there was a strong block effect ($P < 0.0001$), neither the linear contrast of the two RM populations to the original population ($F_{6,2001} = 1.6$, $P = 0.14$) or the interaction of group with block ($F_{48,6982} = 1.2$, $P = 0.20$) was significant, despite very high statistical power. This suggests that effects of selection, mutation and drift on the traits during the experiments were negligible.

The genetic variance/covariance matrices (G) did not differ significantly (likelihood ratio tests²⁰: RM1 versus original, $\chi^2_{21} = 26.4$, $P = 0.19$; RM2 versus original, $\chi^2_{21} = 9.93$, $P = 0.97$; total $N = 1,985$). Common principal components (CPC) analysis¹⁹ of the phenotypic variance/covariance matrices (P) also did not detect significant differences between the RM and the original matrices. Because the CPC method tends to overestimate differences between matrices²¹, and the total sample was very large ($N = 2,021$), the lack of significant difference is strong evidence that the matrices were not altered by the random mating.

The phenotypic and genetic correlation matrices also showed no evidence for declines associated with linkage disequilibrium (Fig. 1 and Table 1): the 95% confidence intervals were broadly overlapping for all correlations, the average changes were 0.07 or less, and the matrices were very highly correlated. Note that although a few of the genetic correlations declined moderately (up to 0.18), the correlations between the lengths of the filaments and corolla tube did not.

Overall, the evidence indicates that floral correlations are caused by pleiotropy, with no evidence for linkage disequilibrium. The means, variances, covariances and correlations for the six floral traits were highly stable over nine generations. My power to detect linkage disequilibrium depends on many factors, including the proportion of the original correlation caused by linkage disequilibrium, the degree of initial disequilibrium among independent loci affecting the traits, and the weighted mean recombination rate among these loci (refs 11, 13, 22, 23 and Fig. 2). My experiment had a high probability to detect declines in phenotypic correlations (Fig. 2, dotted lines) and in genetic correlations with smaller 95% confidence intervals (Fig. 2, lower dashed line), including those between the filaments and corolla tube. For these, linkage disequilibrium would have been detected unless only a very small proportion of the correlation was caused by linkage disequilibrium or if the correlation was caused by loci with harmonic mean recombination rates in the region of 0.01. For the genetic correlations with larger confidence intervals (Fig. 2, upper dashed line denotes the largest), linkage disequilibrium would still have been detectable if more than half of the correlation was caused by linkage disequilibrium and the mean recombination rate was at least 0.2. A mean recombination rate lower than 0.2 is unlikely in wild radish—an outcrosser with nine pairs of chromosomes^{13,23}—unless loci affecting the traits are tightly clustered on a small subset of these chromosomes.

These pleiotropic correlations would be expected to constrain the

Table 1 Comparisons of phenotypic and additive genetic correlation matrices

	Average difference	Maximum decline	Matrix correlation
Phenotypic (average)	0.008	–0.016	0.997
Phenotypic (residuals)	0.011	–0.015	0.995
Additive genetic (component)	–0.069	–0.169	0.984
Additive genetic (breeding value)	–0.040	–0.177	0.939

Comparisons were between the original and random-mated populations. Four different methods of calculating phenotypic and genetic correlation matrices were used (Methods). Absolute values were not used in calculating the average difference, as recommended previously²⁸, because all of the floral correlations are positive and because the direction of the change is of interest.

independent evolution of floral traits in wild radish, particularly the filaments and corolla tube for which the correlations are very high. Studies of populations in species that have diverged naturally or through artificial selection have also not generally detected changes in the *G*-matrix (refs 19, 24, and references therein; but see also ref. 22). These results also have significance for discussions of possible transient changes in the *G*-matrix caused by selection that creates linkage disequilibrium^{22,23}; no evidence of such changes was found in this study. It has been argued that most estimates of the *G*-matrix are made under relaxed selection and therefore do not reflect the *G*-matrix under selection in natural populations²². At least in the population studied here there was no difference between the two, removing one of the many caveats involved in reconstructing past selection.

The finding of correlations that are due to pleiotropy rather than linkage disequilibrium in an obligately outcrossing species is not unexpected on theoretical grounds¹¹. But this expectation is based on several assumptions, including that of weak selection; epistatic or correlational selection can create linkage disequilibrium in outcrossing species^{11,25}. Because there is evidence for correlational selection on the lengths of the filaments and corolla tube in wild radish^{7,16}, the empirical finding of a lack of linkage disequilibrium is significant. The lack of linkage disequilibrium does not mean that selection has not altered the filament–corolla tube correlations; selection can alter correlations caused by pleiotropy as well¹³. Future studies of how correlations respond to correlational selection would improve our understanding of the evolution of integration among functionally related traits. My study provides a direct demonstration of the mechanisms underlying correlations among quantitative traits in a natural population; similar studies on additional traits and species will be crucial to furthering our knowledge of integration and constraint in evolution. □

Methods

Details of mating design

On average, 3% of mothers had no offspring that successfully produced seed in the next generation owing to failures in germination, mortality or unsuccessful pollination; thus, there may have been some selection. Some or all of the lack of germination (the single most common source of failure) was probably caused by environmental stresses on the mothers, especially insect and pathogen attacks, which were frequent in the greenhouse. Evidence for this comes from the pattern of average germination success, which showed no trend over the generations and little variance between the two replicates in a generation, but varied widely from generation to generation. If germination success or resistance or tolerance to enemies was genetically correlated with the floral traits, then there could have been some correlated evolution in the floral traits. These types of genetic correlation are largely unknown, except that there are no significant genetic correlations between germination time and any of the floral traits measured¹⁵. When a total failure of offspring reproduction from one maternal parent occurred, two plants were chosen randomly, one to be used twice as a male and the other twice as a female to maintain each population close to 300 individuals. Because most individuals had two offspring in the next generation, the effective population size in each population was roughly 600, and the inbreeding coefficient increased by only about 1 in 1,200 each generation²⁶.

Analysis details

I compared *G*-matrices using the program *pcrfl*, which is part of the *Quercus* package²⁰. Restricted maximum-likelihood (REML) estimates were made with and without the pair of matrices constrained to be the same, and the fit of the models was compared using two times the difference in log-likelihood, which has a χ^2 distribution. Because the floral traits were measured on all parents and offspring, and *Quercus* takes advantage of all pedigree information, each of these comparisons are based on two-thirds (each RM replicate was tested against the original population separately) of the complete sample of $N = 1,985$ individuals.

Phenotypic variance/covariance matrices (*P*) and phenotypic correlations (Table 1 and Fig. 1) were calculated in two different ways. Both methods were designed to correct for differences in means among blocks and generations that could bias the correlation estimates if all data were simply pooled. In neither case was a significant difference detected between the *P*-matrices, that is, none of the CPC tests was significant at $P < 0.05$ on the basis of 1,000 randomizations.

In the first method, the phenotypic covariance matrices were estimated from residuals of models correcting for differences in means among generations, blocks nested within generations, replicates (for the two RM populations only) and the interactions between them. Therefore, a single test was done comparing the RM to the original populations with $N = 1,317$ and 704, respectively. This total N is slightly greater than the genetic analysis above because some of the parents did not produce flowering offspring and were therefore

eliminated from the genetic analysis. In the second method, I calculated a *P*-matrix for each block and generation separately, and then averaged the corresponding elements across the five matrices for the original population (three parental and two offspring blocks) and across the ten matrices for the random mated populations (5 blocks $\times$ 2 replicates). The CPC tests were done on these averages.

The genetic correlations (Table 1 and Fig. 1) were also calculated in two ways: variance/covariance component correlations were estimated by REML in *Quercus* (see above), and breeding value correlations were estimated as Pearson's product-moment correlations among breeding values calculated using best linear unbiased prediction (BLUP). BLUPs were estimated using Proc Mixed in SAS²⁷.

Received 15 June; accepted 19 August 2002; doi:10.1038/nature01105.

- Maynard Smith, J. et al. Developmental constraints and evolution. *Quart. Rev. Biol.* **60**, 265–287 (1985).
- Deng, H.-W., Haynatzka, V., Spitze, K. & Haynatzki, G. The determination of genetic covariances and prediction of evolutionary trajectories based on a genetic correlation matrix. *Evolution* **53**, 1592–1599 (1999).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Sunderland, Massachusetts, 1998).
- Clark, A. G. in *Genetic Constraints on Adaptive Evolution* (ed. Loeschke, V.) 25–45 (Springer, Berlin, 1987).
- Paterson, A. H. et al. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721–726 (1988).
- Kearsey, M. J. & Farquhar, A. G. L. QTL analysis in plants; where are we now? *Heredity* **80**, 137–142 (1998).
- Morgan, M. T. & Conner, J. K. Using genetic markers to directly estimate male selection gradients. *Evolution* **55**, 272–281 (2001).
- Futuyma, D. J. *Evolutionary Biology* (Sinauer Associates, Sunderland, Massachusetts, 1998).
- Falconer, D. S. & MacKay, T. F. C. *Introduction to Quantitative Genetics* (Longman, Harlow, 1996).
- Mitchell-Olds, T. Pleiotropy causes long-term genetic constraints on life-history evolution in *Brassica rapa*. *Evolution* **50**, 1849–1858 (1996).
- Lande, R. The genetic correlation between characters maintained by selection, linkage and inbreeding. *Genet. Res.* **44**, 309–320 (1984).
- Wright, S. *Evolution and the Genetics of Populations. Vol. 2. The Theory of Gene Frequencies* (Univ. Chicago Press, Chicago, 1969).
- Lande, R. The genetic covariance between characters maintained by pleiotropic mutations. *Genetics* **94**, 203–215 (1980).
- Nuzhdin, S. V., Dilda, C. & Mackay, F. C. The genetic architecture of selection response: Inferences from fine-scale mapping of bristle number quantitative trait loci in *Drosophila melanogaster*. *Genetics* **153**, 1317–1331 (1999).
- Conner, J. K. & Via, S. Patterns of phenotypic and genetic correlations among morphological and life-history traits in wild radish, *Raphanus raphanistrum*. *Evolution* **47**, 704–711 (1993).
- Conner, J. K. Floral evolution in wild radish: the roles of pollinators, natural selection, and genetic correlations among traits. *Int. J. Plant Sci.* **158**, S108–S120 (1997).
- Hill, J. P. & Lord, E. M. Floral development in *Arabidopsis thaliana*: a comparison of the wild type and the homeotic pistillata mutant. *Can. J. Bot.* **67**, 2922–2936 (1989).
- Shabalina, S. A., Yampolsky, L. Y. & Kondrashov, A. S. Rapid decline of fitness in panmictic populations of *Drosophila melanogaster* maintained under relaxed natural selection. *Proc. Natl Acad. Sci. USA* **94**, 13034–13039 (1997).
- Phillips, P. C. & Arnold, S. J. Hierarchical comparison of genetic variance-covariance matrices. I. Using the Flury hierarchy. *Evolution* **53**, 1506–1515 (1999).
- Shaw, R. G. The comparison of quantitative genetic parameters between populations. *Evolution* **45**, 143–151 (1991).
- Houle, D., Mezey, J. & Galpern, P. Interpretation of the results of common principal components analysis. *Evolution* **56**, 433–440 (2002).
- Shaw, F. H., Shaw, R. G., Wilkinson, G. S. & Turelli, M. Changes in genetic variances and covariances: G whiz! *Evolution* **49**, 1260–1267 (1995).
- Bulmer, M. G. Linkage disequilibrium and genetic variability. *Genet. Res.* **23**, 281–289 (1974).
- Roff, D. A. & Mousseau, T. A. Does natural selection alter genetic architecture? An evaluation of quantitative genetic variation among populations of *Allonemobius socius* and *A. fasciatus*. *J. Evol. Biol.* **12**, 361–369 (1999).
- Whitlock, M. C., Phillips, P. C., Moore, F. B. G. & Tonsor, S. J. Multiple fitness peaks and epistasis. *Annu. Rev. Ecol. System.* **26**, 601 (1995).
- Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
- SAS Institute SAS/STAT User's Guide, Version 8 (SAS Institute, Cary, North Carolina, 1999).
- Willis, J. H., Coyne, J. A. & Kirkpatrick, M. Can one predict the evolution of quantitative characters without genetics? *Evolution* **45**, 441–444 (1991).
- Snedecor, G. W. & Cochran, W. G. *Statistical Methods* (Iowa State Univ. Press, Ames, Iowa, 1989).
- Bulmer, M. G. *The Mathematical Theory of Quantitative Genetics* (Oxford Univ. Press, Oxford, 1985).

Acknowledgements I thank C. Haun, T. Hickox, S. Kercher, J. Mudd, C. Stewart and many undergraduates for help in the greenhouse; P. Phillips for help with the CPC analysis and for programs; F. Shaw for analyses with *Quercus*; E. Brodie, B. Charlesworth, J. Fry, D. Houle, R. Lande, M. Lynch, D. Schluter, M. Turelli, S. Via and M. Whitlock for discussion; and A. Agrawal, M. Duffy, J. Fant, J. Kelly, M. Morgan, P. Phillips, H. Sahli, C. Stewart and S. Via for comments on the manuscript. This work was supported by the US National Science Foundation, the University of Illinois and Michigan State University.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to the author (e-mail: conner@kbs.msu.edu).

Mechanisms and circuitry underlying directional selectivity in the retina

Shelley I. Fried^{*†}, Thomas A. Münch^{†‡} & Frank S. Werblin[§]

^{*} Vision Science, [‡] Helen Wills Neuroscience Institute, and [§] Department of Molecular and Cell Biology, University of California Berkeley, 145 LSA, Berkeley, California 94720, USA

[†] These authors contributed equally to this work

In the retina, directionally selective ganglion cells respond with robust spiking to movement in their preferred direction, but show minimal response to movement in the opposite, or null, direction^{1,2}. The mechanisms and circuitry underlying this computation have remained controversial³. Here we show, by isolating the excitatory and inhibitory inputs to directionally selective cells and measuring direct connections between these cells and presynaptic neurons, that a presynaptic interneuron, the starburst amacrine cell, delivers direct inhibition to directionally selective cells. The processes of starburst cells are connected asymmetrically to directionally selective cells: those pointing in the null direction deliver inhibition; those pointing in the preferred direction do not. Starburst cells project inhibition laterally ahead of a stimulus moving in the null direction. In addition, starburst inhibition is itself directionally selective: it is stronger for movement in the null direction. Excitation in response to null direction movement is reduced by an inhibitory signal acting at a site that is presynaptic to the directionally selective cell. The interplay of these components generates reduced excitation and enhanced inhibition in the null direction, thereby ensuring robust directional selectivity.

Directionally selective (DS) movement detection was first characterized in rabbit retina in a series of classical experiments², which suggested that directional selectivity was mediated by an interaction of excitatory and inhibitory signals, whereby inhibition suppressed excitation for movement in the null direction. This role of inhibition was supported by later studies showing that directional selectivity is lost in the presence of antagonists of GABA (γ -aminobutyric acid), an inhibitory transmitter^{4,5}. Other studies have suggested that starburst amacrine cells supply this crucial inhibitory input to DS cells because they co-fasciculate with the processes of DS cells^{6,7} and release GABA^{8–10}.

But the results of some other studies have been contradictory about the role of starburst cells: laser ablation of starburst cells did not eliminate directional selectivity in rabbit¹¹, but eliminating starburst cells with neurotoxin did eliminate it in mouse¹². The site of interaction between excitation and inhibition is also controversial: one report suggests a postsynaptic interaction at the DS cell dendrites¹³, another suggests a presynaptic location¹⁴. Here we have addressed each of these issues to yield a comprehensive description of the synaptic inputs and circuitry that endow the DS cell with directional selectivity.

The excitatory and inhibitory input currents (Fig. 1b) underlying the DS spiking output (Fig. 1a) were measured in on–off DS cells in response to preferred and null direction movement. We isolated these currents by patch clamping at the appropriate reversal potentials (Methods). Preferred direction movement always ($n = 22/22$) elicited larger excitatory currents than did null direction movement (mean 2.5-fold, range 1.1–5.3), whereas the inhibitory currents were always larger for movement in the null direction (mean 1.6-fold, range 1.2–3.7). The same asymmetries were measured when the movement was confined within the receptive field of the DS cell ($n = 10/10$; Fig. 1c). Thus, similar to the spiking output², directional selectivity of the inputs is achieved by a local computation that does not require crossing of the receptive field

boundaries. The asymmetries in magnitude of the input currents suggest that directional selectivity may be formed at sites that are presynaptic to the dendrites of DS cells. Our results show that both inputs are themselves already directional (but see refs 13, 14), but they are delivered to the dendrites of the DS cell where there is presumably further interaction.

A clue to how these asymmetries in magnitude are formed can be found in the spatial profile of the inhibitory input. This profile was offset with respect to the dendritic field of the DS cell. We observed this by correlating the positions at which excitation and inhibition were generated with the dendritic field of the DS cell in response to an edge moving in the preferred direction (Fig. 1d and Methods). The excitatory response profile fell symmetrically around the dendritic field of the DS cell, but the inhibitory response profile was shifted towards the null side of the DS dendritic field. This was not a consequence of the direction of movement: a similar offset in the inhibitory profile was seen for a stimulus moving in the null direction (Fig. 1e). Nor were these profiles a consequence of movement at all: a similar offset of the inhibitory profile could be generated by mapping the receptive field with stationary test flashes moved to different positions along the preferred–null axis of the DS cell receptive field (Fig. 1f).

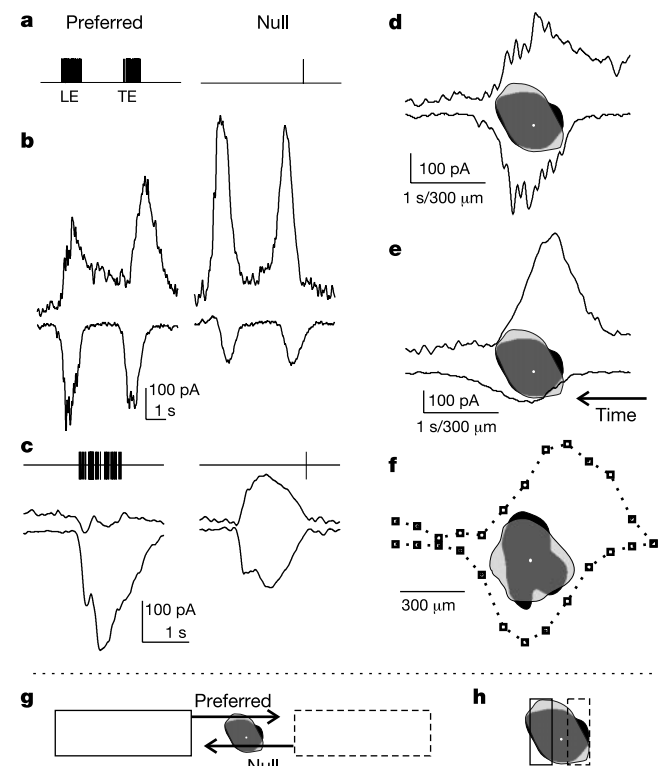


Figure 1 Excitatory and inhibitory inputs to DS ganglion cells: asymmetries in magnitude and space. **a**, Spiking response to movement in the preferred and null directions (the stimulus is shown in **g**). Leading (LE) and trailing (TE) edges of the stimulus generate spiking. **b**, Excitatory (downward deflections) and inhibitory (upward deflections) inputs in response to the same stimulus. **c**, Spiking, inhibition and excitation in response to a stimulus constrained within the dendritic field (depicted in **h**). **d**, **e**, Excitation and inhibition (from **b**) plotted as a function of the stimulus edge position (see Methods) for preferred (LE, **d**) and null direction movements (TE, **e**). Shaded regions indicate off (dark) and on (light) dendrites. Bright spot indicates soma. Current traces in **e** are reversed to reflect right to left movement (indicated by time arrow). **f**, Profiles of excitation and inhibition in response to 100-μm square stationary flashes along the preferred–null axis. **g**, The stimulus bar (300 × 900 μm) used in **a**, **b**, **d** and **e**. Outlines indicate start and end positions. **h**, The stimulus bar (300 × 100 μm) used in **c**. In all figures, axes are aligned so that the preferred direction is from left to right.

These three measurements of excitatory and inhibitory response profiles suggest that excitation is symmetrically arranged around the dendritic field of the DS cell but that there is a region outside the dendritic field of the DS cell on the null side, an 'inhibitory lobe', that is populated by a separate set of neuronal elements that detect movement and project inhibition laterally to the dendrites of the DS cell. A consequence of this inhibitory lobe is that, for null direction movement, the inhibitory signal arrives at the DS cell before excitation (Fig. 1b, e). For the cell shown in Fig. 1f, the inhibitory lobe was offset by about 150 μm (range 50–200 μm , $n = 6$).

We thought that starburst amacrine cells were likely to populate the inhibitory lobe. Their dendritic arbour is broad enough (about 300 μm)¹⁵ to carry signals laterally by more than 150 μm , which was the dimension of the inhibitory offset shown in Fig. 1f. To test whether starburst cells do provide direct inhibition to the DS cells, we simultaneously patched a DS cell and a starburst cell whose soma was located in the inhibitory lobe of the DS cell. Depolarization of such a null-side starburst cell elicited inhibition in the DS cell, confirming that the inhibitory lobe is comprised of starburst cells

($n = 3/3$; Fig. 2b). There were variations in the time course of the inhibitory currents for each cell pair (data not shown), but starburst depolarizations above -20 mV always elicited inhibition. We did not see any excitation from starburst cells in this lobe. We made similar attempts to measure connectivity from starburst cells on the preferred side of the DS cell (Fig. 2a, $n = 3$), but found neither excitatory nor inhibitory activity. Figure 2c summarizes the results of the null-side cell pairs. The steady-state inhibitory currents measured in the DS cell are plotted as a function of starburst cell depolarization for each cell pair. For comparison, a typical I - V curve for a preferred-side cell pair is also shown.

The measurements showing a lack of excitatory input from starburst cells to DS cells were unexpected. It is known that starburst cells release acetylcholine^{16–18}, and that DS cells show cholinergic sensitivity^{4,19}. If the excitatory connectivity exists then it is probably sparse, being expressed by cells other than the ones we measured (starburst cells overlap greatly²⁰), or it is very subtle as compared with the inhibitory inputs. The results of our double patch experiments are, however, consistent with the existence of only an inhibitory lobe on the null side and the lack of any offset lobe on the preferred side of the DS dendrites (Fig. 1f).

Notably, there was an anatomical correlation with our physiological measurements: the cell pairs on the null side showed $17.5 \pm 1.4\%$ (mean $\pm$ s.d.) co-fasciculation (Fig. 2e), whereas those on the preferred side (Fig. 2d) showed only $6.4 \pm 2.9\%$ co-fasciculation ($P < 0.01$, unpaired t -test).

Could directional selectivity originate in the starburst cells themselves? Previous recordings from starburst cells detected no obvious directionality of their light response²¹, but several groups have theorized that starburst cells might release transmitter asymmetrically^{22,23}. The inhibitory input from starburst cells can be measured in isolation in the DS cell by confining the movement of a stimulus within the inhibitory lobe, which lies completely outside the dendritic field of the DS cell in a region where there is no excitatory input (Fig. 2g). Within this lobe, movement in the null direction elicited a larger inhibitory current than did movement in the preferred direction (Fig. 2f). Starburst cells in this lobe contact DS cells with only their left processes. This suggests that starburst cells release more inhibitory transmitter if a stimulus moves outward along these processes.

We draw three conclusions from the results shown in Fig. 2. First, the connectivity between a starburst cell and a DS cell depends on

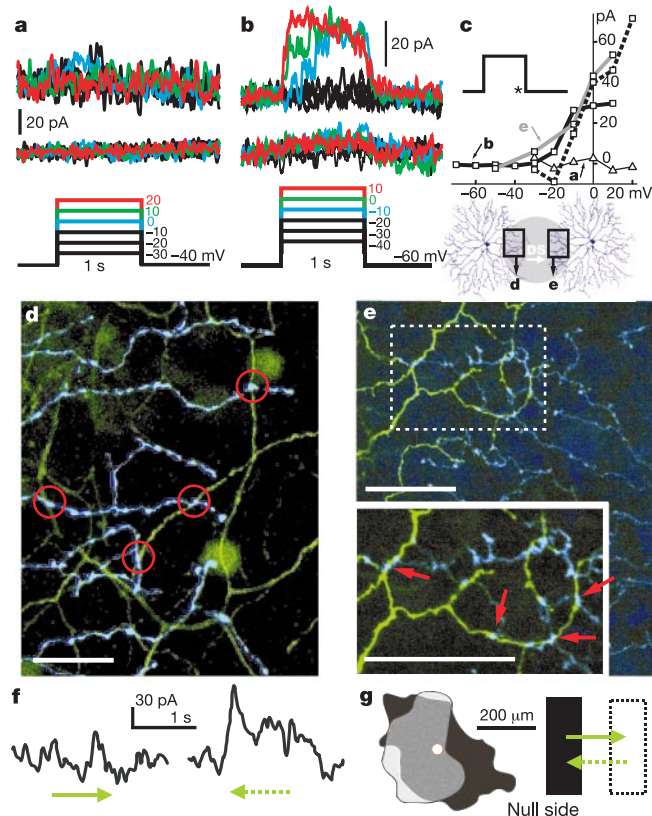


Figure 2 Starburst cells on the null side supply directionally selective inhibition to DS cells. **a, b**, Inhibitory (top traces) and excitatory (middle traces) input currents measured in DS cells when neighbouring starburst cells on the preferred side (**a**) or null side (**b**) were stimulated with depolarizing voltage steps (bottom traces). **c**, Inhibitory input currents measured in DS cells at the indicated time (asterisk) as a function of the stepping voltage of a starburst cell for all null-side (squares) and one preferred-side (triangles) cell pairs. I - V curves from the cells shown in **a, b** and **e** are indicated. **d, e**, Confocal image slices ($\sim 1.5\text{-}\mu\text{m}$ thick) of the contact region between DS (green) and starburst (cyan) pairs. There is co-fasciculation on the null side (**e**, red arrows; inset shows magnified view of outlined region), but mainly crossings on the preferred side (**d**, red circles). **f**, Inhibitory input currents to a DS cell when a stimulus contained within the inhibitory lobe (depicted in **g**) is moved in the preferred (left) or null (right) directions. **g**, The stimulus black bar ($300 \times 100\text{ }\mu\text{m}$) used in **f**. Shaded regions indicate off (dark) and on (light) dendrites. Bright spot indicates soma. Scale bars, 20 μm (**d, e**).

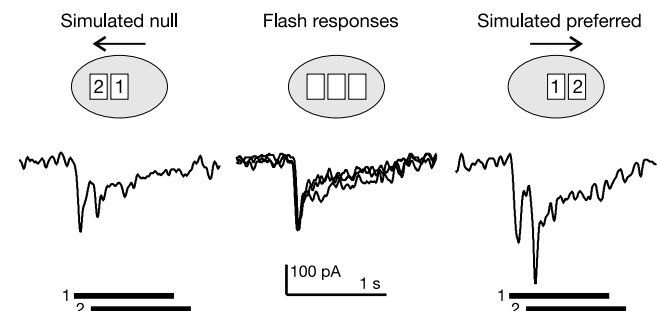


Figure 3 Null movement vetoes excitatory input to the DS cell. Middle, three superimposed excitatory responses to flashed bars ($100 \times 60\text{ }\mu\text{m}$) at three locations (separated by $60\text{ }\mu\text{m}$) within the dendritic field of the cell. Flashes at each of the three locations generate similar responses. Right, excitatory response to two flashes (1 then 2), simulating preferred direction movement. Left, excitatory response to two flashes (1 then 2), simulating null direction movement. Grey ovals indicate the dendritic field of the DS cell. White rectangles indicate the spatial positions of the flashes. Horizontal bars at bottom of the left and right panels indicate the timing sequence of the two flashes (167-ms delay between onsets of 1 and 2).

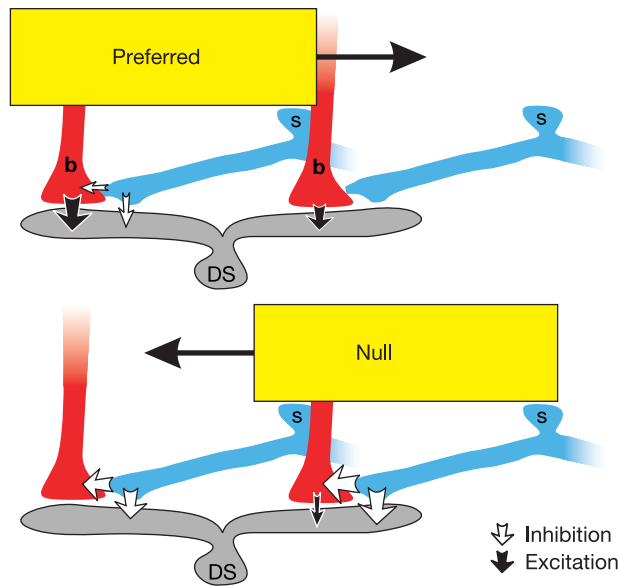


Figure 4 Proposed circuitry underlying the DS response. The processes of starburst cells ('s', blue) that point in the null direction provide inhibition to DS cell dendrites ('DS', grey). Starburst processes respond best to movement away from the cell body, which makes the inhibitory input delivered to DS cells larger for movement in the null direction (bottom panel) than for movement in the preferred direction (top panel). An additional inhibition acts presynaptically to reduce excitation for null direction movement. Although this presynaptic inhibition is depicted as coming from starburst cells, our results do not rule out the existence of another type of cell. The excitatory input to DS cells probably comes from bipolar cells ('b', red) and may also have a cholinergic component from other starburst cells. For movement in the null direction, the inhibitory input reaches each sub-region of the DS cell ahead of the stimulus edge and therefore before excitation. For movement in the preferred direction, inhibition lags behind excitation.

the relative location of the starburst cell: starburst cells on the null side, but not the preferred side, supply inhibition, an idea that was proposed previously in the co-transmission model²⁴. Second, starburst cells detect movement at a site that is displaced from the site where they deliver inhibition to the DS cell, consistent with the proximo-distal segregation of input and output sites²⁵. Third, this inhibition delivered by starburst cells is itself directionally selective, a finding that is complemented by studies showing that a stronger depolarization is measured at the starburst soma in response to centrifugal stimuli (ref. 26 and T.A.M., S.I.F and F.S.W., unpublished data), and that a larger Ca^{2+} signal is measured in starburst cell processes in response to outward motion²⁶.

Could the directional selectivity inherent in the inhibitory input contribute to directional selectivity in excitation? To examine this relationship, we tested the effect of inhibition on excitation by simulating movement with a paired sequence of flashes (Fig. 3). The first flash was presented at the centre of the dendritic field of a DS cell. We measured its effect on a second flash, presented to either side of the first flash. When the second flash was presented to the right, simulating movement in the preferred direction, the excitation elicited by the second flash simply added to that of the first flash. But when the second flash was presented to the left, simulating movement in the null direction, the excitatory response to the second flash was suppressed completely. We quantified this difference by extracting the maximum amplitude of excitation for each recording and then comparing simulated motion to the sum of the two flashes presented individually. The simulated null motion was $47 \pm 19\%$ (mean $\pm$ s.d.) of the sum of the individual flashes, whereas the simulated preferred motion was $97 \pm 33\%$ of the sum of the individual flashes ($n = 6$, data not shown). These results

suggest that inhibition is broadcast only to the left of the stimulus to interfere with excitation for movement in the null direction. This inhibition seems to act presynaptically, modulating the excitatory input before it arrives at the DS cell. It therefore confers directional selectivity to the excitatory input. In the preferred direction, a few cells showed a supralinear summation (data not shown), suggesting a facilitatory mechanism may exist.

Directional selectivity seems to be mediated by complementary mechanisms, as shown in the proposed circuitry in Fig. 4. The starburst cells themselves provide a directionally selective inhibitory input to the DS cell at a site displaced in the null direction from the site of movement. A similarly displaced inhibition, possibly also from starburst cells, suppresses the excitatory signal before excitation arrives at the DS cell, so excitation also seems to be directionally selective. Robust directional selectivity is generated by a combination of increased inhibition and suppressed excitation in the null direction.

Our study moves the unknown issues related to directional selectivity to sites upstream of the DS cells and raises new questions. How do starburst cells come to favour centrifugal motion? Is it a simple biophysical property of the cell, or does it involve synaptic interactions? What is the role of acetylcholine in directional selectivity? Perhaps even more intriguing is the issue of how the spatially asymmetric connectivity between starburst cells and DS cells is established during development. □

Methods

Electrophysiology

We recorded from 40 'on-off' DS ganglion cells and 6 pairs of 'on' starburst amacrine and DS cells in light-adapted whole-mount retinas of 2.5-kg New Zealand white rabbits using EPC-7 amplifiers (Heka). The retina was continuously superfused at $7\text{--}10\text{ ml min}^{-1}$ with Ames solution (Sigma; pH 7.4, 36°C), equilibrated with 95% O_2 and 5% CO_2 . Spiking was recorded with a loose cell-attached electrode ($3\text{--}4\text{ M}\Omega$), filled with Ames solution. For DS cell recordings, whole-cell electrodes ($5\text{--}8\text{ M}\Omega$) were filled with (in mM): 112.5 CsMeSO₄, 1 Mg SO₄, 7.8×10^{-3} CaCl₂, 0.5 BAPTA, 1,2-bis(2-aminophenoxy)ethane-*N,N,N',N'*-tetraacetic acid tetrapotassium salt, 10 HEPES, 4 ATP-Na₂, 0.5 GTP-Na₃, 5 lidocaine *N*-ethyl bromide (QX314-Br), 7.5 neurobiotin chloride (pH 7.2). For starburst cells, we used 1.5 mM EGTA instead of BAPTA, and the concentrations for CaCl₂ and neurobiotin chloride were 0.5 and 6.7 mM, respectively. Before double patch experiments, the tissue was incubated for 10–15 min in $2\text{--}10\text{ }\mu\text{M}$ 4',6-diamidino-2-phenylindole dihydrochloride (DAPI; Molecular Probes) to stain starburst cell bodies²⁷. We measured input currents as described²⁸ by voltage clamping the DS cell at the reversal potential for chloride channels (-60 mV) or ligand-gated nonspecific cation channels (0 mV) to measure excitation or inhibition, respectively. We assessed the effectiveness of space clamp in control experiments (data not shown) by applying SR-95531, a GABA_A antagonist, which blocked the inhibitory signal. Clamping at 0 mV showed no input currents, as expected in the presence of a good space clamp. To plot the $I\text{--}V$ curves (Fig. 2c), we averaged the inhibitory currents measured in the DS cells over a 200 ms window at the end of the depolarizing pulse to the starburst cell. The data acquisition software (RED) was written by M. Wang, T. Lan and D. Handwerker. Data were analysed in Mathematica (Wolfram Research) or Matlab (MathWorks).

Light stimulus

Light stimuli were projected onto the retina as described²⁸. The stimulus for the experiments in Fig. 1a, b, d and e was a bar of $300 \times 900\text{ }\mu\text{m}$ (velocity $300\text{ }\mu\text{m s}^{-1}$). The long axis of the bar was oriented parallel to the preferred-null axis of the cell. The centre of the bar started $900\text{ }\mu\text{m}$ from the soma, moved $1,800\text{ }\mu\text{m}$ in the preferred direction, and then back in the null direction to the original starting point. The stimulus for the experiments in Fig. 1c and 2f was a bar of $100 \times 300\text{ }\mu\text{m}$ (velocity $300\text{ }\mu\text{m s}^{-1}$). The long axis of the bar was oriented perpendicular to the preferred-null axis of the cell. The travel distance of the bar was $200\text{ }\mu\text{m}$. For the experiment in Fig. 1f, $100\text{--}\mu\text{m}$ squares were flashed for 1 s. Thirteen successive flashes were presented $50\text{--}\mu\text{m}$ apart along the preferred-null axis with flash number 7 centred over the soma. At each position the maximum amplitude of excitation and inhibition ('on' response) were determined and used to represent the strength of the input receptive field at that location. The lines in Fig. 1f connect the maxima from each of the 13 flash locations. For the experiment in Fig. 3, bars of $100 \times 60\text{ }\mu\text{m}$ were flashed for 1 s. The long axis was perpendicular to the preferred-null axis of the cell. Flashes were spaced by $60\text{ }\mu\text{m}$. The second flash followed 167 ms after the onset of the first flash. The stimulus contrast was 200% for each experiment.

Fluorescence imaging

Alexa Fluor 488 (Molecular Probes) was added to the intracellular electrode for fluorescent imaging. Cell morphology was obtained by fluorescently imaging the dendrites after electrophysiology had been completed. The image was captured and stored for later analysis using an intensified camera (SIT 66, Dage-MTI) and a video capture board

(PVR2500, Perception). By tracing the dendrites in each sublamina, we obtained the dendritic field limits. The on and off sublamina were imaged by adjusting the depth of focus. In Figs 1 and 2, the extent of the on and off sublaminae are represented by light grey or black outlines, respectively.

Space-time alignment

We obtained Fig. 1d and e by correlating the spatial position of the dendritic field of the DS cell with the time response of the input currents. The position of the stimulus was known at all points in time and could therefore be aligned with the position of the dendrites. The horizontal scale of the figure was adjusted using the stimulus velocity such that the scale bar represents both 300 μm and 1 s. Therefore, any vertical line through the figure would give the position of the leading (Fig. 1d) or trailing (Fig. 1e) edge of the stimulus and the corresponding response magnitudes. The current trace in Fig. 1e is reversed to take into account the right-to-left stimulus movement.

Confocal reconstruction

After the whole-cell patch-clamp recordings, the retinas were fixed, permeabilized and incubated with streptavidin-conjugated Alexa Fluor 488 conjugate to stain the neurobiotin-filled cells. During the double patch experiments, the starburst cell pipette additionally contained Lucifer Yellow (lithium salt, Molecular Probes), which was recognized with a rabbit antibody against Lucifer Yellow (Chemicon International; used at a dilution of 1:100) and visualized with a Rhodamine-Red-conjugated donkey antibody against rabbit IgG (Jackson ImmunoResearch; used at a dilution of 1:200). Images of the whole-mount retina were taken at two different wavelengths (488 nm for ganglion cell morphology, 568 nm to image the starburst cell) on a Zeiss LSM 510 confocal microscope. We determined the relative amount of co-fasciculation between recorded starburst and DS cells by dividing the length of co-fasciculating processes by the total length of starburst dendrites in the region of overlap.

Received 6 June; accepted 2 September 2002; doi:10.1038/nature01179.

- Barlow, H. B. & Hill, R. M. Selective sensitivity to direction of movement in ganglion cells of rabbit retina. *Science* **139**, 412–414 (1963).
- Barlow, H. B. & Levick, W. R. Mechanism of directionally selective units in rabbits retina. *J. Physiol. (Lond.)* **178**, 477–504 (1965).
- Vaney, D. I., He, S., Taylor, W. R. & Levick, W. R. *Motion Vision—Computational, Neural, and Ecological Constraints* (ed. Zeil, J.) 13–56 (Springer, Berlin, 2001).
- Kittila, C. A. & Massey, S. C. Pharmacology of directionally selective ganglion cells in the rabbit retina. *J. Neurophysiol.* **77**, 675–689 (1997).
- Caldwell, J. H., Daw, N. W. & Wyatt, H. J. Effects of picrotoxin and strychnine on rabbit retinal ganglion-cells—lateral interactions for cells with more complex receptive fields. *J. Physiol. (Lond.)* **276**, 277–298 (1978).
- Famiglietti, E. V. Dendritic Costratification of on and on-off directionally selective ganglion-cells with starburst amacrine cells in rabbit retina. *J. Comp. Neurol.* **324**, 322–335 (1992).
- Vaney, D. I. & Pow, D. V. The dendritic architecture of the cholinergic plexus in the rabbit retina: selective labeling by glycine accumulation in the presence of sarcosine. *J. Comp. Neurol.* **421**, 1–13 (2000).
- O'Malley, D. M., Sandell, J. H. & Masland, R. H. Corelease of acetylcholine and GABA by the starburst amacrine cells. *J. Neurosci.* **12**, 1394–1408 (1992).
- Brecha, N., Johnson, D., Peichl, L. & Wässle, H. Cholinergic amacrine cells of the rabbit retina contain glutamate-decarboxylase and γ -aminobutyrate immunoreactivity. *Proc. Natl Acad. Sci. USA* **85**, 6187–6191 (1988).
- Vaney, D. I. & Young, H. M. GABA-like immunoreactivity in cholinergic amacrine cells of the rabbit retina. *Brain Res.* **438**, 369–373 (1988).
- He, S. G. & Masland, R. H. Retinal direction selectivity after targeted laser ablation of starburst amacrine cells. *Nature* **389**, 378–382 (1997).
- Yoshida, K. *et al.* A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron* **30**, 771–780 (2001).
- Taylor, W. R., He, S. Y., Levick, W. R. & Vaney, D. I. Dendritic computation of direction selectivity by retinal ganglion cells. *Science* **289**, 2347–2350 (2000).
- Borg-Graham, L. J. The computation of directional selectivity in the retina occurs presynaptic to the ganglion cell. *Nature Neurosci.* **4**, 176–183 (2001).
- Famiglietti, E. V. Starburst amacrine cells and cholinergic neurons—mirror-symmetric on and off amacrine cells of rabbit retina. *Brain Res.* **261**, 138–144 (1983).
- Masland, R. H., Mills, J. W. & Hayden, S. A. Acetylcholine-synthesizing amacrine cells—identification and selective staining by using autoradiography and fluorescent markers. *Proc. R. Soc. Lond. Ser. B* **223**, 79–100 (1984).
- Masland, R. H. & Mills, J. W. Autoradiographic identification of acetylcholine in the rabbit retina. *J. Cell Biol.* **83**, 159–178 (1979).
- Linn, D. M., Blazynski, C., Redburn, D. A. & Massey, S. C. Acetylcholine-release from the rabbit retina mediated by kainate receptors. *J. Neurosci.* **11**, 111–122 (1991).
- Masland, R. H. & Ames, A. Responses to acetylcholine of ganglion-cells in an isolated mammalian retina. *J. Neurophysiol.* **39**, 1220–1235 (1976).
- Famiglietti, E. V. Starburst amacrine cells—morphological constancy and systematic variation in the anisotropic field of rabbit retinal neurons. *J. Neurosci.* **5**, 562–577 (1985).
- Peters, B. N. & Masland, R. H. Responses to light of starburst amacrine cells. *J. Neurophysiol.* **75**, 469–480 (1996).
- Borg-Graham, L. J. & Grzywacz, N. M. *Single Neuron Computation* (ed. Zornetzer, S. E.) 347–375 (Academic, London, 1992).
- Pozanski, R. R. Modeling the electrotonic structure of starburst amacrine cells in the rabbit retina—a functional interpretation of dendritic morphology. *Bull. Math. Biol.* **54**, 905–928 (1992).
- Vaney, D. I. *Progress in Retinal Research* (ed. Chader, G. J.) 49–100 (Oxford, Pergamon, 1990).
- Famiglietti, E. V. Synaptic organization of starburst amacrine cells in rabbit retina—analysis of serial thin-sections by electron-microscopy and graphic reconstruction. *J. Comp. Neurol.* **309**, 40–70 (1991).
- Euler, T., Detwiler, P. B. & Denk, W. Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature* **418**, 845–852 (2002).

- Mills, S. L. & Massey, S. C. Morphology of bipolar cells labelled by DAPI in the rabbit retina. *J. Comp. Neurol.* **321**, 133–149 (1992).
- Roska, B. & Werblin, F. Vertical interactions across ten parallel, stacked representations in the mammalian retina. *Nature* **410**, 583–587 (2001).

Acknowledgements We thank K. Okazaki for help with confocal preparations and image analysis; J. Hurtado, R. Kramer and C. Kretschmann for discussion; and H. Barlow, R. Froemke, E. Isacoff, X. Ren and R. Zucker for comments on the manuscript. This work was supported by grants from the Office of Naval Research, the National Eye Institute, and an NIH training grant in Vision Science to the University of California Berkeley (S.F.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.W. (e-mail: werblin@socrates.berkeley.edu)

Multiple forms of synaptic plasticity triggered by selective suppression of activity in individual neurons

Juan Burrone, Michael O'Byrne & Venkatesh N. Murthy

Department of Molecular & Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

The rules by which neuronal activity causes long-term modification of synapses in the central nervous system are not fully understood. Whereas competitive or correlation-based rules result in local modification of synapses, homeostatic modifications allow neuron-wide changes in synaptic strength, promoting stability^{1,2}. Experimental investigations of these rules at central nervous system synapses have relied generally on manipulating activity in populations of neurons^{1,3–6}. Here, we investigated the effect of suppressing excitability in single neurons within a network of active hippocampal neurons by overexpressing an inward-rectifier potassium channel. Reducing activity in a neuron before synapse formation leads to a reduction in functional synaptic inputs to that neuron; no such reduction was observed when activity of all neurons was uniformly suppressed. In contrast, suppressing activity in a single neuron after synapses are established results in a homeostatic increase in synaptic input, which restores the activity of the neuron to control levels. Our results highlight the differences between global and selective suppression of activity, as well as those between early and late manipulation of activity.

Activity is thought to play a role in the formation and modification of neural circuits^{3,7}. The specific manner in which activity modifies synaptic strength remains a matter of speculation, and is controversial even in relatively well-studied systems^{8–12}. Two general classes of synaptic modification have emerged as targets of concerted investigation. Hebbian modification has been studied at the level of single neurons or single synapses over a timescale of hours^{13,14}. In contrast, homeostatic modifications have been studied over longer timescales, but have involved uniform changes in activity in entire networks of central neurons^{4,5,15}. The relation between these two rules of synaptic modification has not been studied experimentally.

We sought to determine the consequences of chronic suppression of activity in single neurons within an active network (Fig. 1a). To this end, we used the inward-rectifier potassium channel Kir2.1, which has been used to suppress excitability in mammalian superior cervical ganglion cells and chick hair cells^{16,17}. Overexpression of this channel hyperpolarizes the neuron and decreases its resting

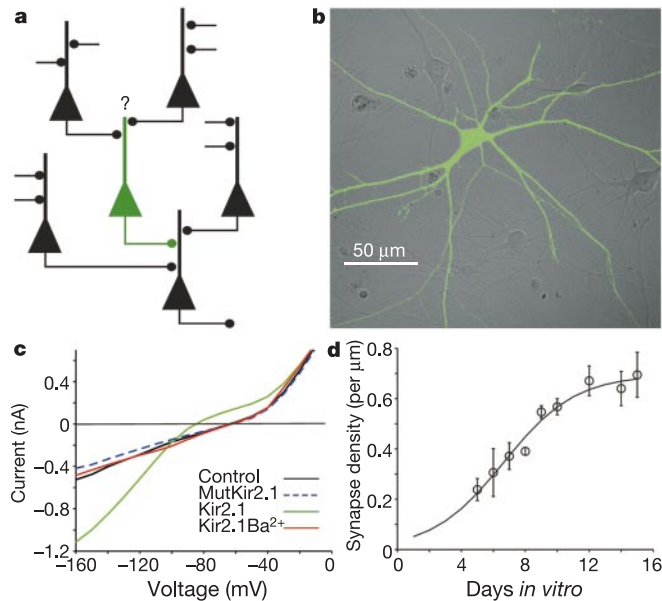


Figure 1 Overexpression of Kir2.1 in individual neurons depresses excitability. **a**, Examination of the long-term consequences of suppressing activity in a single cell (green) embedded in a network of active cells (black). **b**, Superposition of fluorescence and transmitted light images showing the normal morphology of a pyramidal neuron expressing Kir2.1 and EGFP. Transfection was on day 10, and images were obtained on day 14. **c**, Current-voltage relationship showing that cells overexpressing Kir2.1 have significantly larger inward-rectifying currents than mutKir2.1 or untransfected control cells; this additional current was abolished by extracellular Ba²⁺. **d**, Time course of synapse formation in cultures. The number of active presynaptic terminals was measured using the vesicle dye FM4-64, and each point is an average of at least three cells. The continuous line is a sigmoidal function with half-maximum at about 6.5 days.

membrane resistance. Once the spike threshold is reached, however, the Kir2.1 current contributes negligibly to the spike waveform. Therefore, expression of Kir2.1 causes the neuron to be less excitable than neighbouring neurons, leading to a persistent imbalance in activity.

We used a bicistronic construct made from coding sequences of human Kir2.1 and enhanced green fluorescent protein (EGFP) separated by an internal ribosome entry site (IRES) to transfect hippocampal neurons in culture (ref. 16; Fig. 1b). To control for nonspecific effects caused by overexpression of Kir2.1, we carried out parallel experiments using a non-conducting mutant version of Kir2.1 (mutKir2.1) with the pore region amino acid motif GYG mutated to AAA. The low efficiency of transfection using the calcium phosphate method¹⁸ allowed us to minimize interaction between multiple inactivated cells. Neurons overexpressing Kir2.1 exhibited prominent inward-rectifying currents (Fig. 1c), and their resting potential and input resistance were significantly below that of control and mutant Kir2.1-expressing cells (Table 1; $P < 0.01$). These differences were abolished by low concentrations (300 μM) of extracellular Ba²⁺ (Fig. 1c and Table 1). In current-clamp recordings, larger currents were required to bring Kir2.1 neurons to threshold (data not shown). Expression of the additional current

Table 1 Passive parameters

Cell treatment	Resting potential (mV)*	Input resistance (M Ω)
Control	-65 $\pm$ 2.1 (10)	166 $\pm$ 11 (10)
MutKir2.1	-64 $\pm$ 0.6 (10)	216 $\pm$ 12 (10)
Kir2.1	-73 $\pm$ 2.3 (13)	63 $\pm$ 25 (13)
Kir2.1 + Ba ²⁺	-63 $\pm$ 1.4 (13)	219 $\pm$ 14 (13)

The number of cells is indicated in parentheses.

* Corrected for junction potential.

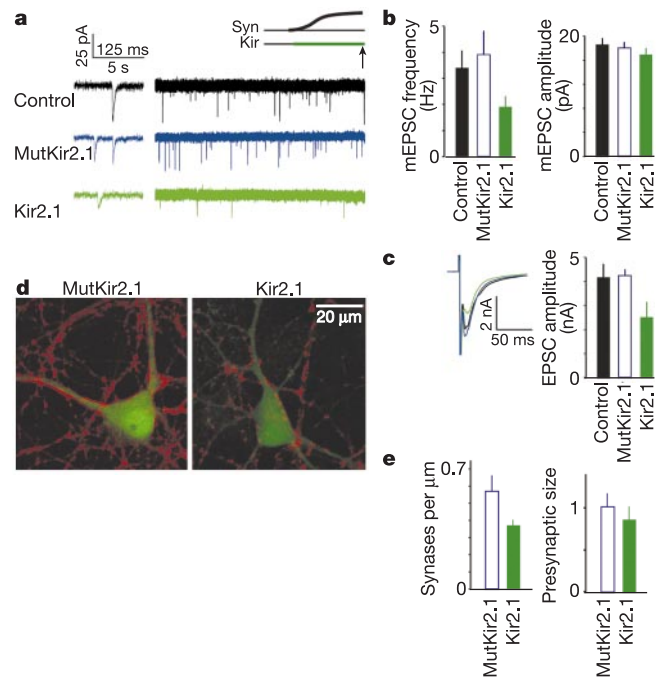


Figure 2 Suppression of activity before synapse formation leads to a reduction in synaptic input. **a**, Examples of recordings from neurons expressing Kir2.1 or mutKir2.1, and an untransfected control neuron on day 14. To the left of each trace, a small segment is shown at expanded timescale to indicate typical miniature excitatory postsynaptic currents (mEPSCs). The top right inset shows the experimental time line with time course of synapse formation (Syn), duration of Kir2.1 overexpression (green) and time of recording (arrow). **b**, The average rate of mEPSCs in cells expressing Kir2.1 is lower than in control or mutKir2.1 cells ($P < 0.005$), but the quantal amplitudes are not different. **c**, Examples of action-potential-evoked EPSCs recorded from single Kir2.1 (green), mutKir2.1 (blue) and control (black) cells. The average EPSC was significantly lower in Kir2.1 cells than in the other two ($P < 0.01$). **d**, Superimposed images of the green (EGFP) and red (FM4-64) fluorescence from a mutKir2.1 cell and Kir2.1 cell taken at the same instrument settings. **e**, The average synapse density for Kir2.1 neurons was significantly smaller than that of mutKir2.1 cells ($P < 0.01$). The average intensity of the individual synapses (normalized), a measure of the presynaptic vesicle pool size, was smaller for Kir2.1 neurons ($P < 0.05$).

reached a stable level within 48 h, and was maintained for several days (see Supplementary Information). Hippocampal networks in dissociated culture normally maintain action-potential firing at low rates owing to their recurrent connections and spontaneous vesicle release. If expression of Kir2.1 leads to decreased excitability, the average firing rates should be lower than normal. As predicted, 24 h after transfection spike-firing rates measured using cell-attached patch-clamp recording were significantly reduced (control, 0.51 ± 0.19 Hz; Kir2.1, 0.08 ± 0.06 Hz, $P < 0.01$, Kolmogorov-Smirnov test, which is used for all comparisons below).

Previous experiments have suggested that synapses can form in the absence of activity in dissociated cultures^{4,19–22}. In those experiments, activity was blocked globally, affecting all neurons in culture. At the vertebrate neuromuscular junction the effect of local manipulation of activity is markedly different from that of global manipulation²³. Therefore, we wondered whether more selective manipulation of activity might uncover mechanisms that are invisible to uniform suppression of activity. To ensure proper timing of suppression of excitability, we first characterized the time course of synapse formation in our cultures by labelling active presynaptic terminals with styryl dyes²⁴. Functional presynaptic terminals could be detected as early as 4 days *in vitro*, but a rapid increase in the number of synapses occurred between day 6 and 10 (Fig. 1d). After day 10, the density of synapses did not increase

Table 2 Synaptic parameters

Cell treatment	mEPSC frequency (Hz)	mEPSC amplitude (pA)	Synapse density (μm^{-1})	Presynaptic pool size (normalized)
Control	3.4 ± 0.65 (21)	18.6 ± 1.02 (21)	—	—
Kir2.1 (early)	1.9 ± 0.41 (37)	16.9 ± 1.07 (37)	0.37 ± 0.034 (8)	0.83 ± 0.14 (8)
Kir2.1 (early) + Ba^{2+}	1.9 ± 0.56 (9)	15.7 ± 1.78 (9)	—	—
MutKir2.1	3.9 ± 0.91 (23)	18.0 ± 0.97 (23)	0.57 ± 0.093 (6)	1.00 ± 0.15 (6)
Kir2.1 (early) + TTX	11.7 ± 4.13 (13)	22.6 ± 3.01 (13)	0.53 ± 0.029 (6)	1.22 ± 0.10 (6)
MutKir2.1 (early) + TTX	—	—	0.59 ± 0.029 (6)	1.15 ± 0.1 (6)
Control + TTX	12.1 ± 3.20 (13)	22.5 ± 2.03 (13)	—	—
Kir2.1 (late)	5.7 ± 1.01 (15)	17.7 ± 1.5 (15)	0.64 ± 0.081 (6)	1.24 ± 0.14 (6)

The number of cells is indicated in parentheses. Early, transfection before synapse formation; late, after synapse formation.

significantly. The rate of miniature excitatory postsynaptic currents (mEPSCs) seems to undergo a similar increase and stabilization²¹.

To examine the effect of reducing excitability of a neuron before synapse formation, we transfected neurons 2–5 days *in vitro*, when synapse density was low. We then used whole-cell patch-clamp recordings to measure the frequency and amplitude of α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor-mediated mEPSCs on day 14 (Fig. 2). Compared with control or mutKir2.1 neurons, Kir2.1-transfected neurons had fewer mEPSCs (Fig. 2a, b; $P < 0.005$) whose average amplitude was not significantly different (Fig. 2b and Table 2; $P > 0.1$). We next tested whether the reduced rate of mEPSCs in cells transfected with Kir2.1 led to smaller evoked synaptic currents. To obtain a measure of the total synaptic input to a single neuron, we used extracellular stimulation to evoke action potentials synchronously in all neurons. Evoked EPSCs were smaller on average in cells expressing Kir2.1 than in control or mutKir2.1 cells (Fig. 2c; Kir2.1, 2.50 ± 0.64 nA; control, 4.16 ± 0.56 nA; mutKir2.1, 4.24 ± 0.26 nA, $P < 0.01$, $n = 6$ each), suggesting a reduced functional input to neurons overexpressing Kir2.1.

To determine whether the decrease in the number of functional synaptic inputs was due to a reduction in the number of synapses, we measured the density of active presynaptic terminals labelled with FM4-64 (Fig. 2d). We found significantly fewer functional presynaptic boutons terminating on Kir2.1 neurons than on mutKir2.1 neurons (Fig. 2e and Table 2; $P < 0.001$). The functional size of the fewer synapses made on Kir2.1 neurons, estimated from the intensity of FM4-64 staining, was smaller than those made on mutKir2.1 neurons (Fig. 2e and Table 2; $P < 0.05$). Therefore, fewer synapses were made onto neurons with reduced excitability, and these synapses were presumably weaker than control synapses because of their smaller presynaptic vesicle pool.

The reduction in synaptic inputs to the Kir2.1-expressing cells might be due to a nonspecific reduction of cell viability or function. Several observations argue against this possibility. First, the use of the pore mutant Kir2.1 helped us rule out effects that are independent of the ionic conduction of the channel. Second, the intrinsic electrical properties of transfected neurons were not altered beyond that expected from the expression of potassium currents—for example, action potentials were intact, and the membrane potential and resting conductance were restored to control levels on addition of extracellular Ba^{2+} . Third, the axons of transfected neurons had normal presynaptic terminals, as judged by the expression of vesicle-associated membrane protein (VAMP) fused to EGFP²⁵ (see Supplementary Information). The fourth and most compelling argument against nonspecific effects in transfected neurons is that we could prevent the reduction in synaptic inputs by silencing the entire network of neurons using tetrodotoxin (TTX; see below).

Does the reduction in synaptic inputs to the less active neuron result from a relative reduction in activity compared with neighbouring active neurons, or is it due to an absolute reduction in activity? To distinguish between these possibilities, we blocked sodium action potentials in the entire network using TTX for 5–9 days before recordings were done (Fig. 3a). Although the transfected

neuron will continue to have a lower resting potential than its neighbours, none of the neurons will produce action potentials. We found that within days of the global activity blockade, transfected neurons regained synaptic inputs and resembled their untransfected neighbours (Fig. 3a). The rates of mEPSCs increased uniformly across all neurons (Fig. 3b; $P < 0.001$), consistent with our previous findings that presynaptic vesicle pools are larger in TTX-treated cultures²⁶. The amplitudes of mEPSCs were also larger (Fig. 3b; $P < 0.01$), as seen for cortical neurons⁵. The number of functional presynaptic boutons terminating on Kir2.1 neurons was indistinguishable from those on mutKir2.1 neurons (Fig. 3c, d). Furthermore, TTX treatment increased the size of the total recycling vesicle pool in boutons terminating on both Kir2.1 and mutKir2.1 neurons (Fig. 3d; see also ref. 26). Although it is possible that blocking spike activity using TTX activates an entirely different cellular programme, these results nevertheless indicate that it is not simply the lack of activity that results in fewer inputs to transfected neurons. One hypothesis is that the decrease in synaptic input observed in Kir2.1 neurons without TTX is triggered by the difference in activity levels between different neurons.

Previous studies in dissociated cultures have found that global block of activity using TTX leads to an increase in quantal amplitude⁵, presynaptic release probability²⁶, spontaneous release²⁷

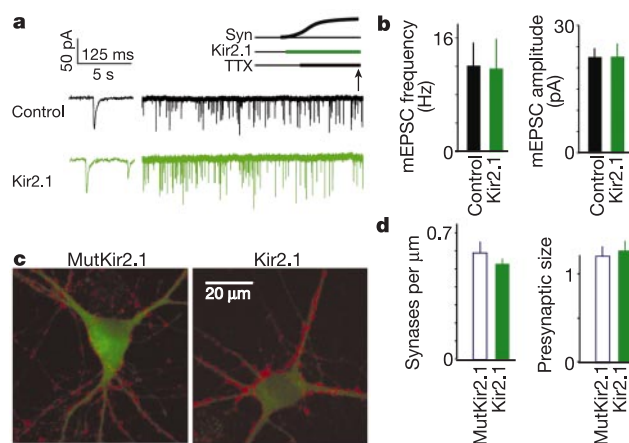


Figure 3 Blocking activity uniformly using TTX equalizes synaptic inputs to Kir2.1 and control cells. **a**, Sample recordings from a neuron transfected with Kir2.1 and a control neuron, both grown in TTX for 3 days. The experimental time line is shown in the top right inset. **b**, The frequency of mEPSCs for cells expressing Kir2.1 and for control neurons increase substantially after TTX treatment, but no difference was observed between Kir2.1 and control neurons in TTX ($P > 0.5$). Similarly, the average amplitudes of mEPSCs increased after TTX treatment ($P < 0.005$), but no difference was observed between Kir2.1 and control cells after TTX treatment ($P > 0.5$). **c**, Sample images illustrating that cells transfected with Kir2.1 gain active presynaptic boutons after TTX treatment. **d**, The density of presynaptic terminals on Kir2.1 cells was similar to those on mutKir2.1 neurons ($P > 0.5$). The average presynaptic vesicle pool size was similar ($P > 0.2$) for Kir2.1 and mutKir2.1 cell synapses after TTX treatment, but both had larger vesicle pools compared with non-TTX cells ($P < 0.01$).

and the ultrastructurally determined size of the synapse²⁶. At the *Drosophila* neuromuscular junction, overexpression of Kir2.1 in muscle triggers a homeostatic increase in transmitter release²⁸. We wondered what effect silencing a single neuron would have on established rather than developing synapses, and how this might compare with the global blockade of activity using TTX. To investigate this, we transfected neurons with Kir2.1 at 10 days *in vitro*, when synapse number is close to steady state (Fig. 1d). We confirmed that expression of Kir2.1 led to a reduction in the resting membrane potential and membrane resistance. In contrast to suppression of excitability before synapse formation, we found that late suppression led to a significant increase in the frequency of mEPSCs (Fig. 4a, b; $P < 0.005$), whereas the quantal amplitude was not different (Fig. 4b; $P > 0.1$). The increased rate of mEPSCs in cells transfected with Kir2.1 was accompanied by an increase in the total evoked synaptic currents (Fig. 4c; Kir2.1, 1.76 ± 0.4 nA; control, 1.38 ± 0.35 nA; $n = 6$, $P < 0.05$), suggesting a larger functional input to neurons overexpressing Kir2.1. The modest increase in evoked release might be sufficient to overcome the effect of Kir2.1, which inactivates on depolarization. The effects of the global activity block on established synapses were larger than those caused by Kir2.1 (Fig. 3), perhaps reflecting the stronger block of activity by TTX compared with Kir2.1.

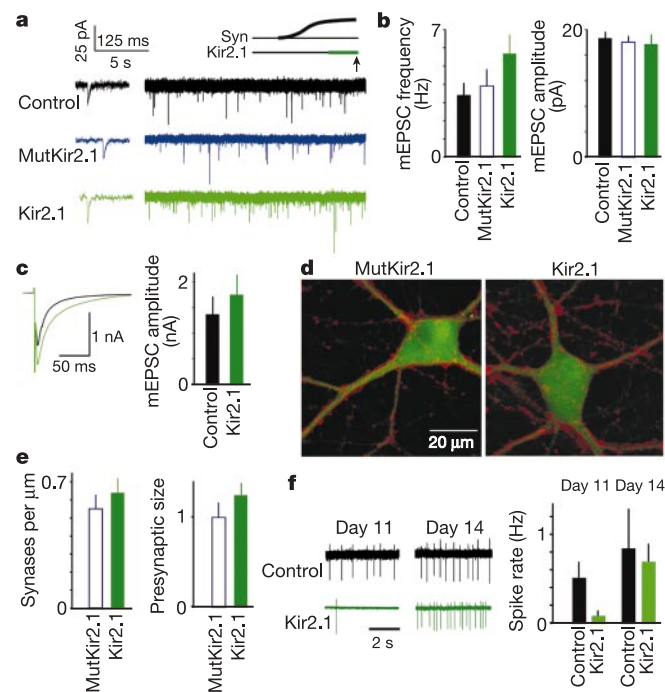


Figure 4 Suppression of excitability after synapse formation leads to a homeostatic increase in synaptic inputs. **a**, Example recordings of mEPSCs from neurons transfected with Kir2.1 or mutKir2.1. **b**, The average rate of mEPSCs in cells transfected with Kir2.1 was higher than those of control or mutKir2.1 neurons ($P < 0.005$), but the amplitudes were not different ($P > 0.2$). **c**, Evoked EPSCs recorded from a neuron transfected with Kir2.1 (green) and from a control neuron (black). The average EPSC was larger in Kir2.1 neurons ($P < 0.01$). **d**, Examples showing the similarity in the number of presynaptic boutons terminating on Kir2.1-transfected and mutKir2.1 neurons. **e**, The average density of presynaptic terminals was not significantly different between mutKir2.1 neurons and neurons expressing Kir2.1 ($P > 0.5$), but the intensity of individual boutons was larger for Kir2.1 neurons ($P < 0.05$). **f**, Increase in mEPSC rate restores firing rates of Kir2.1 neurons. One day after transfection, spike frequencies were significantly lower in Kir2.1 cells compared with control cells ($P < 0.01$). By day 14 (4 days after transfection) however, spike frequencies were similar in transfected and control cells ($P > 0.3$). Sample recordings of spikes are shown (left).

The number of presynaptic boutons terminating on Kir2.1 neurons was not significantly different from mutKir2.1 neurons (Fig. 4d, e). However, the total recycling pool of vesicles was larger in synapses terminating on Kir2.1 neurons (Fig. 4e; $P < 0.01$). Although the larger vesicle pool could explain the higher frequency of mEPSCs, we have not ruled out additional contributions from an increase in the number of synapses with detectable AMPA currents^{29,30}. We next tested whether the increase in functional synaptic inputs could compensate for the decrease in intrinsic excitability of the Kir2.1 cells, and restore normal spike activity. One day after transfection, the spike firing rates in Kir2.1 cells were substantially reduced (Fig. 4f). Notably, over the next 3 days, the spike-firing rate returned to control levels (Fig. 4f; control, 0.84 ± 0.45 Hz; Kir2.1, 0.70 ± 0.21 Hz; $n = 8$, $P > 0.5$), presumably due to a compensatory increase in transmitter release. This normalization of firing rate was not observed in neurons whose activity was suppressed before synapse formation (mutKir2.1, 0.82 ± 0.23 Hz; Kir2.1, 0.30 ± 0.18 Hz; $n = 8$, $P < 0.01$ measured at day 9).

Using a relatively simple and accessible model system, we have uncovered a role for activity in the regulation of synaptic inputs to a neuron. Silencing a neuron before synapses are formed leads to a reduction in synaptic inputs to this less-excitabile neuron, implying that axons, when faced with a choice of active or silent neurons, prefer to make synapses on more easily excited neurons. It remains to be determined whether the loss of synaptic inputs is due to a lack of synapse formation itself, or the inability to stabilize nascent synapses. In contrast to the effect of early reduction of excitability, we found that suppressing excitation after synapses are already in place led to a homeostatic increase in synaptic input strength. Although the mechanism underlying this increase might be similar to that seen after global block of activity^{1,5,6,15}, our experiments establish that the homeostatic increase is cell autonomous. Also, we find that the increase in synaptic inputs is quite precise and restores the frequency of action potential firing to control levels; such normalization of activity is not possible with TTX treatment. These results complement findings at the *Drosophila* neuromuscular junction, where overexpression of Kir2.1 in muscle fibres resulted in an increase in quantal content, without changes in quantal amplitude or synapse morphology²⁸.

Our experiments indicate that decreased neuronal activity can have multiple and widely different effects depending on the developmental stage of neuronal networks. The experimental system described here might enable the study of the mechanisms underlying the differential effects of early and late suppression of activity, as well as activity dependent, long-term synaptic plasticity in general. □

Methods

Culture and transfection

Neonatal hippocampal neurons (P0–P2) were cultured on glia feeder layers as described previously²⁶. Density of neurons at time of analysis (14–15 days) was approximately 50 neurons per mm². Cells were transfected with the bicistronic EGFP–IRES–Kir2.1 construct¹⁶ using the calcium phosphate transfection protocol¹⁸. As a control, we used channels with mutations in the pore residues, GYG to AAA, made using the Quikchange XL site-directed mutagenesis kit (Stratagene). The concentration of plasmid DNA was optimized empirically to produce robust transfection in a small number of individual neurons on each coverslip. Neurons were transfected on 2–5 days *in vitro*, or 10–11 days *in vitro*.

Electrophysiology

Patch-clamp recordings from cultured hippocampal neurons were performed with an Axopatch 200B amplifier. Electrodes, whose tip resistances were 4–7 MΩ, contained (in mM): 130 potassium gluconate, 10 NaCl, 1 EGTA, 0.133 CaCl₂, 2 MgCl₂, 10 HEPES, 3.5 MgATP, and 1 NaGTP. Extracellular perfusion medium contained (in mM): 136 NaCl, 2.5 KCl, 10 HEPES, 10 D-glucose, 2 CaCl₂, and 1.3 MgCl₂. Recordings were filtered at 2–5 kHz and were acquired at 5–10 kHz. For recording mEPSCs and current–voltage relationship, extracellular medium was supplemented with 50 μM picrotoxin and 0.5 or 1 μM TTX. Analysis of mEPSCs was performed using the MiniAnalysis program (Synaptosoft). To determine the effect of increased membrane permeability caused by Kir2.1 overexpression on mEPSC frequency and amplitude, in some experiments we compared recordings made

in the presence and absence of 300 μM extracellular Ba^{2+} . Current–voltage relations for transfected and control neurons were calculated by recording whole-cell currents under voltage clamp. Voltage steps in 10-mV increments were applied every 1 s. Evoked synaptic currents were recorded using pipettes whose internal solution was supplemented with QX-314 (4 mM), and in the presence of 50–100 μM picrotoxin in the extracellular medium to isolate excitatory synapses. Brief 1-ms pulses were applied to the entire field of neurons using platinum wires separated by about 5 mm. The applied voltage was systematically varied for each recording until the response amplitude became saturated, indicating reliable activation of all axons innervating the recorded neuron. Responses were measured at a holding potential of -70 mV. To record action-potential firing, without perturbing the intracellular environment, we used the cell-attached patch-clamp method. Pipettes similar to those for whole-cell recordings were used, and tight seals were obtained without break-in. Current recordings (under voltage clamp at -70 mV) allowed unambiguous discrimination of spikes with high signal-to-noise ratios.

FM4-64 imaging

To identify functional presynaptic terminals, we labelled recycling synaptic vesicles using 10 μM FM4-64 (Molecular Probes). Neurons were depolarized for 60 s using hyperkalaemic solution (in mM): 78.5 NaCl, 60 KCl, 10 HEPES, 10 D-glucose, 2 CaCl_2 , 1.3 MgCl_2 , 0.05 AP5 (D(-)-2-amino-5-phosphonovaleric acid), 0.005 CNQX (6-cyano-7-nitroquinoxaline-2,3-dione), and 0.001 TTX. Coverslips were then washed in regular extracellular medium without FM4-64 for 10 min before imaging to reduce the background fluorescence caused by non-internalized dye binding to the cell membrane. This protocol has been shown to provide an estimate of the total recycling pool of vesicles. Cells were then imaged in dye-free buffer containing blockers. Earlier experiments also included a de-staining step to release dye from vesicles, and images after de-staining were subtracted from the initial image. As the initial fluorescence and releasable fluorescence were strongly correlated, in most experiments we skipped the de-staining step. Image stacks (Z steps of 1 μm , 7–10 steps) were obtained using a confocal microscope (Olympus Fluoview attached to a BX50WI, $\times 40$, 0.8NA water lens). EGFP and FM4-64 signals were acquired simultaneously using 488-nm excitation, and 510–550-nm band pass and 585 long-pass emission filters, respectively. Transmitted light images were taken separately to identify cell bodies and processes of non-transfected neurons. We chose to measure the density of presynaptic terminals in the proximal regions of the dendrites (about 100 μm from the soma) as these could be identified unequivocally by the EGFP fluorescence. Image analysis was performed in a blind manner with respect to Kir2.1 and *mutKir2.1* neurons. Individual puncta were identified manually after thresholding the images using a value that was 2 standard deviations above the background fluorescence. Analysis of fluorescence intensities was performed with custom-written routines in the MATLAB software environment²⁶.

Analysis

The values for all variables reported were estimated for each cell and averaged across all cells in each group. Errors are reported as standard error of the mean. The Kolmogorov–Smirnov test was used for all statistical comparisons.

Received 16 July; accepted 23 October 2002; doi:10.1038/nature01242.

- Turrigiano, G. G. & Nelson, S. B. Hebb and homeostasis in neuronal plasticity. *Curr. Opin. Neurobiol.* **10**, 358–364 (2000).
- Abbott, L. F. & Nelson, S. B. Synaptic plasticity: taming the beast. *Nature Neurosci.* **3** Suppl., 1178–1183 (2000).
- Katz, L. C. & Shatz, C. J. Synaptic activity and the construction of cortical circuits. *Science* **274**, 1133–1138 (1996).
- Rao, A. & Craig, A. M. Activity regulates the synaptic localization of the NMDA receptor in hippocampal neurons. *Neuron* **19**, 801–812 (1997).
- Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* **391**, 892–896 (1998).
- Liao, D., Zhang, X., O'Brien, R., Ehlers, M. D. & Haganir, R. L. Regulation of morphological postsynaptic silent synapses in developing hippocampal neurons. *Nature Neurosci.* **2**, 37–43 (1999).
- Sanes, J. R. & Lichtman, J. W. Development of the vertebrate neuromuscular junction. *Ann. Rev. Neurosci.* **22**, 389–442 (1999).
- Crowley, J. C. & Katz, L. C. Development of ocular dominance columns in the absence of retinal input. *Nature Neurosci.* **2**, 1125–1130 (1999).
- Hubener, M. & Bonhoeffer, T. Eyes wide shut. *Nature Neurosci.* **2**, 1043–1045 (1999).
- Lin, D. M. *et al.* Formation of precise connections in the olfactory bulb occurs in the absence of odorant-evoked neuronal activity. *Neuron* **26**, 69–80 (2000).
- Zheng, C., Feinstein, P., Bozza, T., Rodriguez, I. & Mombaerts, P. Peripheral olfactory projections are differentially affected in mice deficient in a cyclic nucleotide-gated channel subunit. *Neuron* **26**, 81–91 (2000).
- Zhao, H. & Reed, R. X inactivation of the OCNC1 channel gene reveals a role for activity-dependent competition in the olfactory system. *Cell* **104**, 651–660 (2001).
- Luscher, C., Nicoll, R. A., Malenka, R. C. & Muller, D. Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature Neurosci.* **3**, 545–550 (2000).
- Bi, G. Q. & Poo, M. M. Synaptic modification by correlated activity: Hebb's postulate revisited. *Ann. Rev. Neurosci.* **24**, 139–166 (2001).
- Davis, G. W. & Bezprozvanny, I. Maintaining the stability of neural function: a homeostatic hypothesis. *Annu. Rev. Physiol.* **63**, 847–869 (2001).
- Johns, D. C., Marx, R., Mains, R. E., O'Rourke, B. & Marban, E. Inducible genetic suppression of neuronal excitability. *J. Neurosci.* **19**, 1691–1697 (1999).
- Holt, J. R. *et al.* Functional expression of exogenous proteins in mammalian sensory hair cells infected with adenoviral vectors. *J. Neurophysiol.* **81**, 1881–1888 (1999).
- Xia, Z., Dudek, H., Miranti, C. K. & Greenberg, M. E. Calcium influx via the NMDA receptor induces immediate early gene transcription by a MAP kinase/ERK-dependent mechanism. *J. Neurosci.* **16**, 5425–5436 (1996).

- Craig, A. M., Blackstone, C. D., Haganir, R. L. & Banker, G. Selective clustering of glutamate and GABA receptors opposite synaptic terminals releasing the corresponding neurotransmitters. *Proc. Natl Acad. Sci. USA* **91**, 12373–12377 (1994).
- O'Brien, R. J. *et al.* The development of excitatory synapses in cultured spinal neurons. *J. Neurosci.* **17**, 7339–7350 (1997).
- Gomperts, S. N., Carroll, R., Malenka, R. C. & Nicoll, R. A. Distinct roles for ionotropic and metabotropic glutamate receptors in the maturation of excitatory synapses. *J. Neurosci.* **20**, 2229–2237 (2000).
- Cottrell, J. R., Dube, G. R., Egles, C. & Liu, G. Distribution, density, and clustering of functional glutamate receptors before and after synaptogenesis in hippocampal neurons. *J. Neurophysiol.* **84**, 1573–1587 (2000).
- Balace-Gordon, R. J. & Lichtman, J. W. Long-term synapse loss induced by focal blockade of postsynaptic receptors. *Nature* **372**, 519–524 (1994).
- Betz, W. J. & Bewick, G. S. Optical analysis of synaptic vesicle recycling at the frog neuromuscular junction. *Science* **255**, 200–203 (1992).
- Li, Z. & Murthy, V. N. Visualizing post-endocytic traffic of synaptic vesicles at hippocampal synapses. *Neuron* **31**, 593–605 (2001).
- Murthy, V. N., Schikorski, T., Stevens, C. F. & Zhu, Y. Inactivity produces increases in neurotransmitter release and synapse size. *Neuron* **32**, 673–682 (2001).
- Bacci, A. *et al.* Chronic blockade of glutamate receptors enhances presynaptic release and downregulates the interaction between synaptophysin-synaptobrevin-vesicle-associated membrane protein 2. *J. Neurosci.* **21**, 6588–6596 (2001).
- Paradis, S., Sweeney, S. T. & Davis, G. W. Homeostatic control of presynaptic release is triggered by postsynaptic membrane depolarization. *Neuron* **30**, 737–749 (2001).
- Liao, D., Hessler, N. A. & Malinow, R. Activation of postsynaptically silent synapses during pairing-induced LTP in CA1 region of hippocampal slice. *Nature* **375**, 400–404 (1995).
- Isaac, J. T. R., Nicoll, R. A. & Malenka, R. C. Evidence for silent synapses: implications for the expression of LTP. *Neuron* **15**, 427–434 (1995).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank E. Marban for the gift of the EGFP–IRES–Kir2.1 construct. We also thank the members of our laboratory for discussion. This work was supported by grants from the National Institutes of Health and the NSF. V.N.M. is a Sloan Foundation Fellow, a Pew Scholar, an ERLB Foundation Scholar and a National Alliance for Research on Schizophrenia and Depression (NARSAD) Young Investigator. J.B. is a Grable Investigator of NARSAD.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.N.M. (e-mail: vnmurthy@fas.harvard.edu).

Functional improvement of dystrophic muscle by myostatin blockade

Sasha Bogdanovich^{*†}, Thomas O. B. Krag^{*†}, Elisabeth R. Barton^{*}, Linda D. Morris^{*}, Lisa-Anne Whitemore[‡], Rexford S. Ahima[§] & Tejvir S. Khurana^{*}

^{*} Department of Physiology and Pennsylvania Muscle Institute, University of Pennsylvania School of Medicine, 3700 Hamilton Walk, Richards A-601, Philadelphia, Pennsylvania 19104-6085, USA

[‡] Musculoskeletal Sciences Department, Wyeth Research, Cambridge, Massachusetts 02140, USA

[§] Division of Endocrinology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

[†] These authors contributed equally to this work

Mice^{1,2} and cattle³ with mutations in the myostatin (GDF8) gene show a marked increase in body weight and muscle mass, indicating that this new member of the TGF- β superfamily is a negative regulator of skeletal muscle growth. Inhibition of the myostatin gene product is predicted to increase muscle mass and improve the disease phenotype in a variety of primary and secondary myopathies. We tested the ability of inhibition of myostatin *in vivo* to ameliorate the dystrophic phenotype in the *mdx* mouse model of Duchenne muscular dystrophy (DMD)^{4–8}.

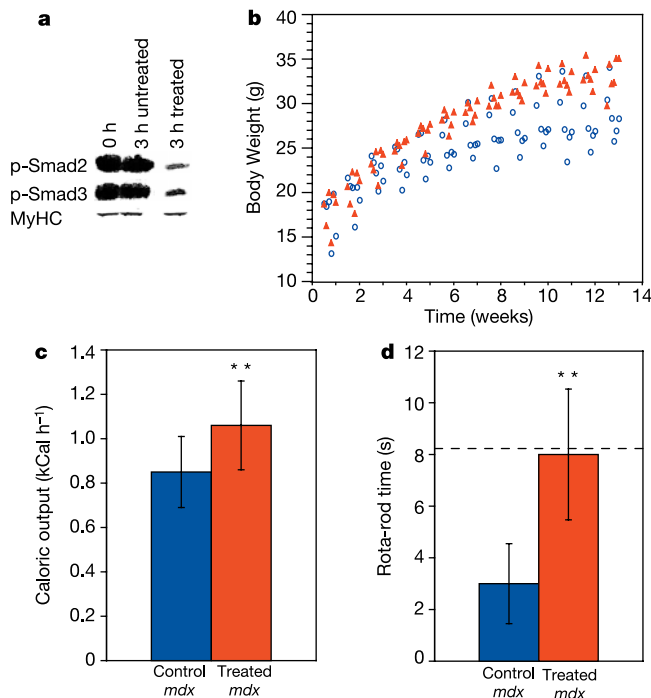


Figure 1 Metabolic consequences of myostatin blockade *in vivo* in *mdx* mice. **a**, Treated C₂C₁₂ cells express less phospho-Smad 2 and 3 than PBS treated controls (lower panel shows MyHC control for even loading and transfer). **b–d**, Comparison of body weights (**b**), caloric output (**c**) and endurance time (**d**) on a Rota-Rod apparatus between treated *mdx* mice and control *mdx* mice. Treated mice (red) had significantly increased body weight (33.11 ± 2.08 versus 28.99 ± 2.94 g; $n = 12$; $P < 0.03$), caloric output (1.06 ± 0.20 versus 0.85 ± 0.16 kCal h⁻¹; $n = 4$; $P < 0.01$) and endurance time (8 ± 2.5 versus 3 ± 1.5 s; $n = 6$; $P < 0.002$) than controls (blue).

Blockade of endogenous myostatin by using intraperitoneal injections of blocking antibodies for three months resulted in an increase in body weight, muscle mass, muscle size and absolute muscle strength in *mdx* mouse muscle along with a significant decrease in muscle degeneration and concentrations of serum creatine kinase. The functional improvement of dystrophic muscle by myostatin blockade provides a novel, pharmacological strategy for treatment of diseases associated with muscle wasting such as DMD, and circumvents the major problems associated with conventional gene therapy in these disorders.

Cultured C₂C₁₂ muscle cells were treated with blocking antibodies generated against myostatin^{9,10} to investigate the signalling mechanism of myostatin in muscle. Treated cells had decreased concentrations of phospho-Smad 2 and 3 compared with controls (Fig. 1a). Alterations in Smad 2 and 3 phosphorylation would be predicted to regulate the formation of Smad complexes, their nuclear translocation and transcriptional regulation of genes associated with commitment of muscle progenitors^{11,12}.

To test myostatin blockage *in vivo*, 4-week-old male *mdx* mice were treated with weekly intraperitoneal injections of blocking antibodies (dose 60 mg kg⁻¹; treated *mdx* group), and vehicle alone (control *mdx* group) for 3 months. Animals were weighed weekly and growth curves were plotted. Treated mice gained weight faster than controls and weighed significantly more than controls after 3 months of treatment (Fig. 1b), which is consistent with the predicted biological effect of myostatin blockade *in vivo*.

To estimate energy expenditure in live animals, we performed indirect calorimetry¹³. Treated mice had a greater caloric output than controls (Fig. 1c), which is consistent with an increase in muscle mass and body size due to myostatin blockade. To determine whether the increase was functional, we used a Rota-rod to assess whole body muscle strength. *Mdx* mice have previously been shown to have an impaired ability to maintain grip and suspend themselves against gravity on the Rota-rod¹⁴. Treated mice performed better than controls (Fig. 1d), which is consistent with increased functional muscle mass and intact neuromuscular coordination *in vivo*.

To quantify the increase of muscle mass, animals were killed and extensor digitorum longus (EDL) muscles were dissected out and weighed. EDL muscles from treated mice weighed significantly more than those from controls (Fig. 2a). Interestingly, the degree of gain of muscle mass was greater than the degree of increase in body weight (Fig. 2b), suggesting that myostatin blockade had a greater proportional effect on muscle than on other organ systems of *mdx* mice. Weights of other muscles including gastrocnemius, tibialis anterior and quadriceps were similarly increased (data not shown). We quantified functional improvement by analysing physiological properties of muscle (Table 1). Treated mice had significantly increased force production during twitch and tetanic contraction (Fig. 2c–f). This increase in muscle strength was proportional to the degree of increase in muscle mass and offers physiological evidence for a functional improvement in *mdx* muscle produced by myostatin blockade *in vivo*.

To determine whether the increase in muscle mass and strength occurred because of hypertrophy or hyperplasia we performed morphometric examination of the EDL muscle (Table 1). Significant increases were noted in whole-muscle cross-sectional area (CSA) and single-fibre area of EDL from the treated mice, indicating true hypertrophy at the single-myofibre level. Frequency histograms of single-fibre areas revealed an overall shift of distribution towards larger areas (Fig. 2g). No significant difference was noted in the number of muscle fibres, total number of nuclei or the nuclei/fibre ratio in either group, suggesting that the increase in muscle mass and size occurred as a result of hypertrophy and not hyperplasia, as has been noted in dominant-negative myostatin mice². A slight increase was noted in the number of centrally nucleated fibres (CNF) in treated mice, indicating fibres that had undergone either regeneration and/or changes in the state of progenitor cell commitment^{8,15}. Because CNF are considered to be more resistant to necrosis¹⁶, their greater proportion might be mechanistic in the improvement in muscle function noted in this study. Interestingly, increased regeneration and/or myogenesis in extraocular muscles has recently been suggested as an important mechanism for the

Table 1 Comparison of physiological and morphometric properties of EDL muscle

	C57BL/10	Control <i>mdx</i>	Treated <i>mdx</i>
CSA (mm ²)	1.4 ± 0.6 (5)	1.5 ± 0.5 (12)	2.0 ± 0.4 (12)**
Absolute force (mN)	356.3 ± 126.8 (10)	370.7 ± 66.5 (12)	491.2 ± 56.5 (12)**
Specific force (mN mm ⁻²)	152.6 ± 58.0 (10)	138.1 ± 30.5 (12)	141.5 ± 22.9 (12)
ECC force decrease (1–5) (%)	n.d.	40.0 ± 11.9 (7)	39 ± 4.6 (7)
CNF (%)	3.5 ± 1.3 (1,554)	41.6 ± 10.3 (4,661)	53.4 ± 3.8 (4,711)**
Single-fibre area (μm ²)	n.d.	1,055.0 ± 578.6 (4,661)	1,321.9 ± 785.1 (4,711)**
Number of myofibres	n.d.	786.2 ± 167.0 (6)	777.8 ± 124.1 (6)
Number of nuclei	n.d.	2,266 ± 339.6 (6)	2158 ± 1,058.3 (6)

Results are presented as means ± s.d.; numbers in parentheses are *n*; asterisks, statistical significance ($P \leq 0.05$). CSA, cross-sectional area; ECC, Eccentric contraction; CNF, centrally nucleated fibres; n.d., not determined.

clinical and pathological sparing of this muscle group in DMD^{17,18}.

To examine for histological evidence of improvement, we sectioned the diaphragm in addition to EDL, because the diaphragm rather than EDL shows degeneration and fibrosis by 16 weeks¹⁹. Histological analysis revealed insufficient foci of degeneration in EDLs of either group to allow comment on whether improvement had occurred (data not shown). However, diaphragm from treated mice (Fig. 3b) showed a decrease in degenerative changes and cellular infiltration compared with controls (Fig. 3a). The inset in Fig. 3b shows diaphragm from normal C57BL/10 mice for comparison. Immunohistochemistry (Fig. 3c) revealed normal concentrations of utrophin enrichment at the synaptic regions of the muscle, suggesting that myostatin blockade ameliorated the dystrophic phenotype via a utrophin-independent mechanism^{20,21}. No differences in utrophin expression were detected by immunoblotting either (data not shown). To ascertain improvement in the pathological status of the whole skeletal musculature, we analysed serum creatine kinase (CK). Elevated CK concentrations are consistently noted with dystrophin deficiency in *mdx* mice and humans owing to sarcolemmal damage^{4,22}. At the start of the trial both

groups had marked elevations of serum CK compared with normal C57BL/10 mice. However, after 3 months of myostatin blockade *in vivo* there was a marked decline in serum CK concentrations in treated mice (Fig. 3d), to almost normal. These decreases in muscle degeneration and serum CK offer histological and biochemical evidence for a functional improvement in *mdx* muscle produced by myostatin blockade *in vivo*.

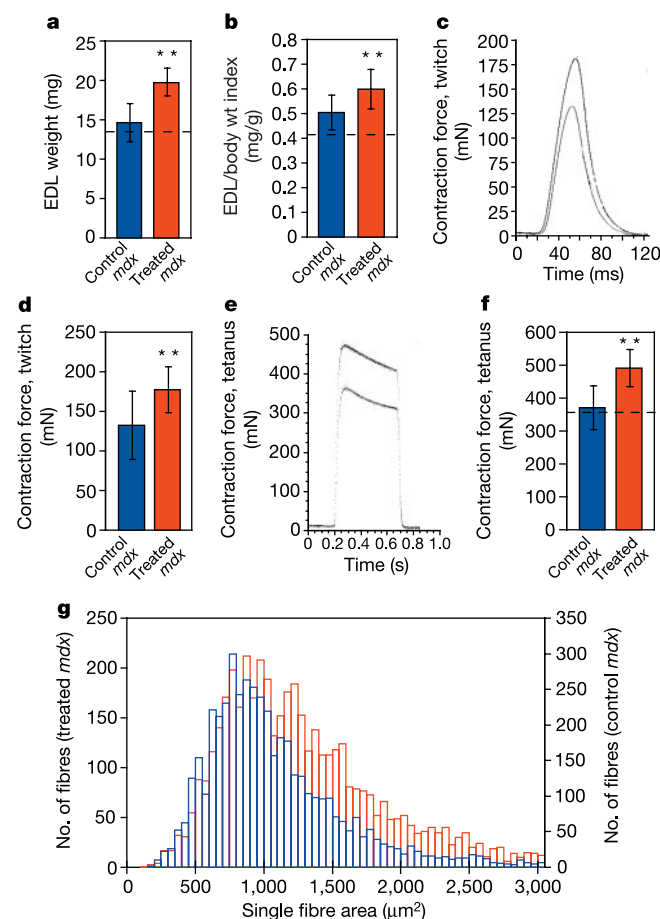


Figure 2 Increase in muscle mass and strength by myostatin blockade. Comparisons of EDL weight (a), EDL/body weight index (b), twitch force (c, d), tetanic force (e, f) and single-fibre areas (g), between treated *mdx* mice (red) and control *mdx* mice (blue). Treated mice had significantly increased EDL weight (19.72 ± 1.76 versus 14.63 ± 2.42 mg; $n = 12$; $P < 0.0001$), had an increased EDL/body weight index (0.6 ± 0.08 versus 0.5 ± 0.07 ; $n = 12$; $P < 0.014$), generated greater twitch force (c, representative traces; d; 177.32 ± 28.95 versus 132.38 ± 43.07 mN; $n = 12$; $P < 0.03$) and tetanic force (e, representative traces; f; 491.23 ± 56.54 versus 370.74 ± 66.47 mN; $n = 12$; $P < 0.003$), and had larger single-fibre area distribution (g), than controls.

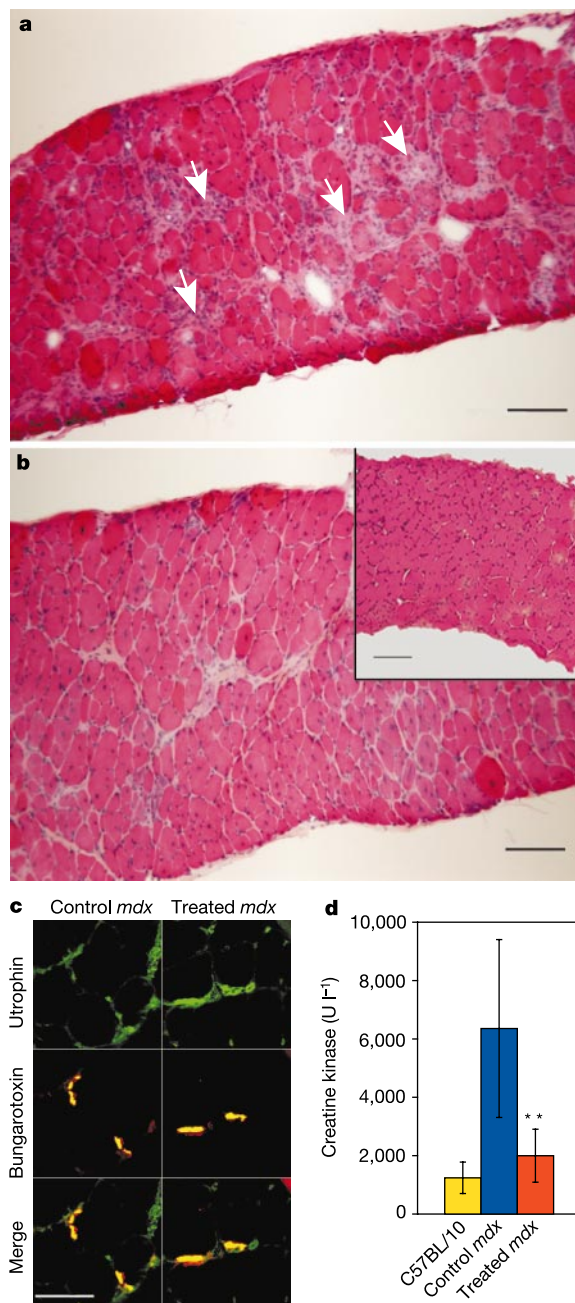


Figure 3 Decreased muscle damage resulting from myostatin blockade. a, b, Control *mdx* mice (a) had significantly greater pathological changes (arrowheads) in the diaphragm than treated mice (b), as evidenced by a lack of cellular infiltration and fibrosis. Inset, normal C57BL/10 mice. Haematoxylin–eosin staining; scale bar, $100 \mu\text{m}$. c, Treated mice (right panels) show no increase in utrophin compared with controls (left panels). Synapses revealed by bungarotoxin staining. Scale bar, $50 \mu\text{m}$. d, Treated mice (red) had significantly decreased CK concentrations compared with controls (blue; 1997.3 ± 907.37 versus $6356.9 \pm 3046.57 \text{ U l}^{-1}$; $n = 6$; $P < 0.005$). The decrease approached C57BL/10 concentrations (yellow; $1241.07 \pm 539.77 \text{ U l}^{-1}$; $n = 4$).

Here we have demonstrated that myostatin blockade *in vivo* in *mdx* mice achieved by pharmacological means resulted in a functional improvement of the dystrophic phenotype, by anatomical, physiological and biochemical criteria. The regimen did not completely reverse dystrophic changes; physiological parameters evaluated by *ex vivo* tests, such as susceptibility to damage by lengthening contractions^{23,24} were not improved. This might be related to initiating myostatin blockade 1 month after birth rather than at birth, an inadequate dosage, or an inherent limitation of the strategy itself. The nature of the improvement indicates that myostatin blockade might be beneficial for a variety of primary and secondary myopathies such as the muscular dystrophies, ageing and muscle loss due to chronic infections or immobilization. Additionally, it might prove to be a useful adjuvant for the management of systemic metabolic disorders such as obesity and diabetes mellitus²⁵. In comparison with conventional transgenic, cell and gene therapy approaches^{26–28}, the current study used a pharmacological approach (administration of a blocking antibody) to improve the dystrophic phenotype functionally. Delivery of blocking antibodies by simple parenteral injections circumvents the need for the generation of specialized viral vectors to deliver gene products. Furthermore, our approach should obviate the problems concerning toxicity or immune response against the vector, which has been of concern in recent therapeutic trials using conventional gene therapy^{29,30}. □

Methods

Physiological studies on isolated muscle

Physiological properties of muscle were analysed with intact *ex vivo* muscle from 16-week-old male *mdx* (C57BL/10ScSn-DMD^{mdx}/J) mice, as described previously^{23,24}. In brief, muscle length was adjusted to achieve maximal twitch response, and this length (L_0) was measured with vernier calipers. Eccentric contraction (ECC) force decrease was calculated from the difference of isometric force generation during the first and fifth tetanus of the standard ECC protocol (supramaximal stimulus of 700 ms, 500 ms isometric phase, 200 ms eccentric phase; total lengthening $L_0/10$; lengthening velocity $0.5 L_0 s^{-1}$). At the end of physiological studies, muscles were flash-frozen in isopentane cooled in liquid nitrogen, and stored at $-80^\circ C$ before being sectioned.

Antibodies and histology

The mouse anti-myostatin monoclonal blocking antibody (clone JA16; Wyeth Research/Genetics Institute) was generated against recombinant myostatin and inhibits the binding of myostatin to its receptor ActRIIB with a half-maximal inhibitory concentration of 500 nM (refs 9, 10, 12). Serial frozen sections (7–12 μm thick) were cut at mid-belly of muscle. Sections were examined after haematoxylin–eosin staining or labelling with anti-laminin antibodies and DNA-binding dye Hoechst 33825 for quantifying nuclear content and other morphological indices. Scion Image 4.02 software (<http://www.scioncorp.com>) was used for morphometric measurements. All cross-sectional myofibres (9,372) and nuclei (26,551) contained in one EDL per animal that was physiologically evaluated in this study were imaged, scored and measured to calculate histological parameters. Affinity-purified BH 11 antibody against utrophin^{20,21} (1 $\mu g ml^{-1}$), rhodamine-conjugated bungarotoxin (4 $\mu g ml^{-1}$; Molecular Probes) and Alexa Fluor® 488 secondary antibodies (Molecular Probes) were used for immunohistochemistry.

Immunoblotting

Cultured C₂C₁₂ cells were treated with anti-myostatin antibody (11.5 $\mu l ml^{-1}$) for 3 h and controls were treated with PBS. Cells were harvested in sample buffer, resolved by SDS–polyacrylamide gel electrophoresis and electrotransferred to poly(vinylidene difluoride) membrane. Efficiency of transfer and the even loading of lanes was verified by using post-transfer Ponceau S staining of the membrane. Membranes were probed with antibodies specific against phospho-Smad 2 and 3 (Santa Cruz Biotechnology) and antibody complexes detected using enhanced chemiluminescence (Pierce).

Biochemical and functional assessment of muscle

Serum CK was measured with the indirect CK colorimetric assay kit and standards (Sigma). The Rota-rod apparatus was designed and built in the laboratory (details available from the authors on request). The apparatus consisted of a hollow plastic rod 6 mm in diameter attached horizontally to a direct-current motor. Rotation speed was calibrated to 18 r.p.m. by using a potentiometer and stopwatch. Indirect calorimetry was performed with the OxyMax Equalflow system (Columbus Instruments). Mice were acclimatized to the test chamber on a 12-h light–dark cycle (light on at 0700h) for 2 days, and night-time energy expenditure was measured at 15-min intervals. Settings used per cycle were as follows: airflow, 500 ml min⁻¹; sample flow, 400 ml min⁻¹; settle time, 120 s; measuring time, 60 s; temperature, 22°C. The caloric output (kCal h⁻¹) was calculated as $3.815 + 1.232 \times \text{respiratory exchange ratio (RER)}$. RER is the ratio between carbon dioxide generated and oxygen consumption.

Statistical analysis

Student's *t*-test was used for determining statistical significance throughout the study. Graphical representation of data uses the following convention: mean $\pm$ s.d.; treated *mdx* mice in red; control *mdx* mice in blue; dashed line represents data from age-matched C57BL/10 normal mice.

Received 10 July; accepted 16 September 2002; doi:10.1038/nature01154.

- McPherron, A. C., Lawler, A. M. & Lee, S. J. Regulation of skeletal muscle mass in mice by a new TGF- β superfamily member. *Nature* **387**, 83–90 (1997).
- Zhu, X., Hadhazy, M., Wehling, M., Tidball, J. G. & McNally, E. M. Dominant negative myostatin produces hypertrophy without hyperplasia in muscle. *FEBS Lett.* **474**, 71–75 (2000).
- Grobet, L. *et al.* A deletion in the bovine myostatin gene causes the double-muscling phenotype in cattle. *Nature Genet.* **17**, 71–74 (1997).
- Bulfield, G., Siller, W. G., Wight, P. A. & Moore, K. J. X chromosome-linked muscular dystrophy (*mdx*) in the mouse. *Proc. Natl Acad. Sci. USA* **81**, 1189–1192 (1984).
- Torres, L. F. & Duchon, L. W. The mutant *mdx*: inherited myopathy in the mouse. Morphological studies of nerves, muscles and end-plates. *Brain* **110**, 269–299 (1987).
- Hoffman, E. P., Monaco, A. P., Feener, C. C. & Kunkel, L. M. Conservation of the Duchenne muscular dystrophy gene in mice and humans. *Science* **238**, 347–350 (1987).
- Hoffman, E. P., Brown, R. H. Jr & Kunkel, L. M. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell* **51**, 919–928 (1987).
- Coulton, G. R., Morgan, J. E., Partridge, T. A. & Sloper, J. C. The *mdx* mouse skeletal muscle myopathy. I. A histological, morphometric and biochemical investigation. *Neuropathol. Appl. Neurobiol.* **14**, 53–70 (1988).
- Thies, R. S. *et al.* GDF-8 propeptide binds to GDF-8 and antagonizes biological activity by inhibiting GDF-8 receptor binding. *Growth Factors* **18**, 251–259 (2001).
- Jalenak, M. *et al.* Transcriptional changes arising from *in vivo* inhibition of myostatin. *J. Neurol. Sci.* **199** (suppl.), S83 (2002).
- Moustakas, A., Souchelnytskyi, S. & Heldin, C. H. Smad regulation in TGF- β signal transduction. *J. Cell Sci.* **114**, 4359–4369 (2001).
- Lee, S. J. & McPherron, A. C. Regulation of myostatin activity and muscle growth. *Proc. Natl Acad. Sci. USA* **98**, 9306–9311 (2001).
- Guyton, A. C. *Textbook of Medical Physiology* (W. B. Saunders, New York, 1995).
- Muntoni, F., Mateddu, A., Marchei, F., Clerk, A. & Serra, G. Muscular weakness in the *mdx* mouse. *J. Neurol. Sci.* **120**, 71–77 (1993).
- Brown, S. C. & Lucy, J. A. *Dystrophin: Gene, Protein and Cell Biology* (Cambridge Univ. Press, Cambridge, 1997).
- Matsuda, R., Nishikawa, A. & Tanaka, H. Visualization of dystrophic muscle fibers in *mdx* mouse by vital staining with Evans blue: evidence of apoptosis in dystrophin-deficient muscle. *J. Biochem. (Tokyo)* **118**, 959–964 (1995).
- McLoon, L. K. & Wirtschauer, J. D. Continuous myonuclear addition to single extraocular myofibers in uninjured adult rabbits. *Muscle Nerve* **25**, 348–358 (2002).
- Fischer, M. D. *et al.* Expression profiling reveals metabolic and structural components of extraocular muscles. *Physiol. Genom.* **9**, 71–84 (2002).
- Stedman, H. H. *et al.* The *mdx* mouse diaphragm reproduces the degenerative changes of Duchenne muscular dystrophy. *Nature* **352**, 536–539 (1991).
- Khurana, T. S., Hoffman, E. P. & Kunkel, L. M. Identification of a chromosome 6-encoded dystrophin-related protein. *J. Biol. Chem.* **265**, 16717–16720 (1990).
- Khurana, T. S. *et al.* Immunolocalization and developmental expression of dystrophin related protein in skeletal muscle. *Neuromuscul. Disord.* **1**, 185–194 (1991).
- Engel, A. G. & Franzini-Armstrong, C. *Myology* (McGraw-Hill, New York, 1994).
- Petrof, B. J., Shrager, J. B., Stedman, H. H., Kelly, A. M. & Sweeney, H. L. Dystrophin protects the sarcolemma from stresses developed during muscle contraction. *Proc. Natl Acad. Sci. USA* **90**, 3710–3714 (1993).
- Moens, P., Baatsen, P. H. & Marechal, G. Increased susceptibility of EDL muscles from *mdx* mice to damage induced by contractions with stretch. *J. Muscle Res. Cell Motil.* **14**, 446–451 (1993).
- Ahima, R. S. & Flier, J. S. Leptin. *Annu. Rev. Physiol.* **62**, 413–437 (2000).
- Tinsley, J. M. *et al.* Amelioration of the dystrophic phenotype of *mdx* mice using a truncated utrophin transgene. *Nature* **384**, 349–353 (1996).
- Gussoni, E. *et al.* Dystrophin expression in the *mdx* mouse restored by stem cell transplantation. *Nature* **401**, 390–394 (1999).
- Ebihara, S. *et al.* Differential effects of dystrophin and utrophin gene transfer in immunocompetent muscular dystrophy (*mdx*) mice. *Physiol. Genom.* **3**, 133–144 (2000).
- Tripathy, S. K., Black, H. B., Goldwasser, E. & Leiden, J. M. Immune responses to transgene-encoded proteins limit the stability of gene expression after injection of replication-defective adenovirus vectors. *Nature Med.* **2**, 545–550 (1996).
- Lozier, J. N. *et al.* Toxicity of a first-generation adenoviral vector in rhesus macaques. *Hum. Gene Ther.* **13**, 113–124 (2002).

Acknowledgements We thank L. Sweeney, C. Bönnemann, J. Khurana (University of Pennsylvania) and L. Edvinsson (Glostrup Hospital, University of Copenhagen and Lund) for guidance and helpful discussions, and N. Wolfman (Wyeth Research/Genetics Institute) for providing reagents and helpful discussions. T.S.K. was a paid consultant for Wyeth Research/Genetics Institute during the study. The study was supported in part by a grant from Wyeth Research/Genetics Institute and a predoctoral fellowship to T.O.B.K. from the Faculty of Medicine, Copenhagen University, Denmark.

Competing interests statement The authors declare competing financial interests: details accompany the paper on *Nature's* website (► <http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to T.S.K. (e-mail: tsk@mail.med.upenn.edu).

Shoot control of root development and nodulation is mediated by a receptor-like kinase

Lene Krusell*, Lene H. Madsen*, Shusei Sato†, Grégoire Aubert‡, Aratz Genua*, Krzysztof Szczygłowski§, Gérard Duc‡, Takakazu Kaneko†, Satoshi Tabata†, Frans de Bruijn||, Eloisa Pajuelo*, Niels Sandal* & Jens Stougaard*

*Laboratory of Gene Expression, Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10, 8000 Aarhus C, Denmark

†Kazusa DNA Research Institute, Kisarazu, Chiba, 292-0812, Japan

‡Institut National de la Recherche Agronomique (INRA), Unité de Recherche de Genétique et d'Amélioration des Plantes (URGAP), BP 86510, 21065 Dijon, Cédex, France

§Agriculture and Agri-Food Canada, SCPFR, 1391 Sandford Street, London, Ontario N5V 4T3, Canada

||Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, B.P. 27 31326 Castanet-Tolosan, Cedex, France

In legumes, root nodule organogenesis is activated in response to morphogenic lipochitin oligosaccharides that are synthesized by bacteria, commonly known as rhizobia¹. Successful symbiotic interaction results in the formation of highly specialized organs called root nodules, which provide a unique environment for symbiotic nitrogen fixation. In wild-type plants the number of nodules is regulated by a signalling mechanism integrating environmental and developmental cues to arrest most rhizobial infections within the susceptible zone of the root^{2–7}. Furthermore, a feedback mechanism controls the temporal and spatial susceptibility to infection of the root system. This mechanism is referred to as autoregulation of nodulation, as earlier nodulation events inhibit nodulation of younger root tissues^{3,4,8}. *Lotus japonicus* plants homozygous for a mutation in the *hypernodulation aberrant root* (*har1*) locus escape this regulation and form an excessive number of nodules^{9–11}. Here we report the molecular cloning and expression analysis of the *HAR1* gene and the pea orthologue, *Pisum sativum*, *SYM29*. *HAR1* encodes a putative serine/threonine receptor kinase, which is required for shoot-controlled regulation of root growth, nodule number, and for nitrate sensitivity of symbiotic development.

To understand the autoregulatory mechanism of nodulation at the molecular level, we have isolated and characterized three mutant alleles from the model legume *L. japonicus*: *har1-1*, *har1-2* and *har1-3* (refs 9, 10). Plants mutated in the *HAR1* locus assume a phenotype characterized by excessive growth of either nodules (hypernodulation; symbiotic phenotype) or lateral roots (bushy root; non-symbiotic phenotype observed in the absence of rhizobia) (see Supplementary Information). The symbiotic phenotype of the *har1* mutants is further characterized by an unusual hypernodulation response (HNR), defined by marked and rapid inhibition of root elongation and shoot growth¹¹.

Split root and grafting experiments in hypernodulating soybean and pea have identified leaf tissue as a principal source of the systemic signal(s) contributing to autoregulation of nodule number^{12–14}. To determine whether the *L. japonicus* symbiotic phenotype is also controlled by the shoot, we performed reciprocal grafting between wild-type and *har1-3* plants. When grown in the presence of rhizobia, *har1-3* shoots grafted onto wild-type roots induced hypernodulation (3.4 ± 1.2 nodules per cm per root; Fig. 1a, d) and the HNR response. The corresponding grafts of wild-type shoots onto *har1-3* mutant roots restored wild-type symbiotic development in the *har1-3* mutant roots (1 ± 0.4 nodules per cm; Fig. 1b, c). Wild type to wild type and *har1-3* to *har1-3* control grafts

resulted in 0.8 ± 0.4 and 4.1 ± 1.5 nodules per cm, respectively. The strong negative effect of *har1* mutations on the main root of mutant plants grown in the absence of rhizobia allowed us to investigate the role of the *HAR1* locus in non-symbiotic root growth. Grafting of wild-type shoots onto *har1-3* roots restores normal root elongation and reduces lateral root formation, resulting in a wild-type phenotype. Together, the results of grafting experiments indicate an important role for the *HAR1* locus in long distance shoot regulation of both nodule organogenesis and root development.

The *HAR1* locus is positioned to the lower arm of *L. japonicus* chromosome 3 (ref. 15) between the amplified-fragment length polymorphism (AFLP) markers E34M61-10F and E35M44-12F (Fig. 2a). The positional cloning strategy for *HAR1* and the physical map of the region is outlined in Fig. 2b, c (see also Methods). In brief, closely linked markers were used to assemble a contig of transformation-competent artificial chromosome (TAC) and λ clones that were sequenced as part of the *L. japonicus* genome-sequencing programme^{16,17}. Subsequent fine mapping located *HAR1* to a 0.2-centimorgan (cM) region (22 kilobases (kb)) delimited by recombination events (Fig. 2b). Sequencing of the complete 22-kb region in the *har1-3* mutant⁹ identified only one change; that is, a deletion removing six amino acids from one of the receptor kinase genes present in this region (Fig. 3a, b). The restriction-fragment length polymorphism (RFLP) resulting from the absence of an *Asp700I* restriction site removed by the deletion is shown in Fig. 3c. Sequencing of the same receptor kinase gene in the *har1-1* and *har1-2* mutants¹⁰ identified point mutations leading to premature stop codons (Fig. 3a, b). Together, physical and genetic map location and identification of mutations in three independent alleles unequivocally identifies the putative receptor kinase gene as *HAR1*.

A partial complementary DNA corresponding to the *HAR1* receptor kinase gene was isolated from a leaf cDNA library derived from wild-type *L. japonicus* Gifu plants, and extended to full length by the 5' rapid amplification of cDNA ends (RACE) procedure. Alignment of the cDNA with the genomic sequence defined a gene structure with one intron, as outlined in Fig. 3a. Sequencing of ten independent 5' RACE products determined the transcription initiation point and full-length cDNA of 3,298 nucleotides. An open reading frame of 2,958 nucleotides is preceded by a leader sequence of 99 nucleotides; the 3' untranslated region is 241 nucleotides. With a DNA sequence identity of 53.2%, the

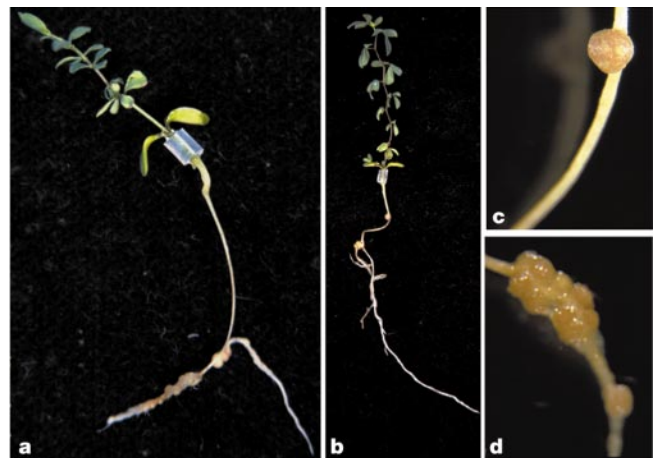


Figure 1 Grafting of *har1* mutants and wild-type plants. **a**, Grafting of mutant shoot to wild-type root results in a mutant root phenotype. **b**, Grafting of wild-type shoot to mutant root restores the wild-type root phenotype. **c**, Close-up view of a nodule on a mutant root grafted to a wild-type shoot. **d**, Close-up of nodules on a wild-type root grafted to a mutant shoot.

L. japonicus *HAR1* gene is similar to the *Arabidopsis* *CLAVATA1* (*CLV1*) receptor kinase gene¹⁸. Furthermore, the position of the intron is conserved between *HAR1* and *CLV1*.

The conceptual *HAR1* protein consists of 986 amino acids with a predicted relative molecular mass of 108,500 (M_r 108.5K) (Fig. 3b). Comparative analysis indicated that *HAR1* belongs to the transmembrane leucine-rich repeat (LRR) receptor-like kinase family including the *Arabidopsis* *BR1*, *ERECTA* and *CLV1* kinases¹⁹. At the amino-terminus, a hydrophobic stretch of 25 amino acids is predicted to act as a signal peptide (Fig. 3b). The signal peptide is followed by a LRR composed of a X ϕ XXNXLS/TGX ϕ PXX ϕ XX ϕ XXLXXL motif (where ϕ and X indicate a hydrophobic and a variable residue, respectively) repeated 21.5 times and flanked by the characteristic paired cysteines²⁰. The presumed intracellular kinase domain of *HAR1* has all the motifs associated with functional serine/threonine kinases¹⁹. Between the LRR and the kinase domain, the algorithms predict a membrane-spanning region flanked by charged residues. From this analysis a topology with an extracellular LRR receiver domain and an intracellular kinase is suggested. The stop codons in *har1-1* and *har1-2* are predicted to generate truncated translation products of 675 amino acids and 918 amino acids that lack the entire carboxy-terminal kinase or the kinase domains X and XI, respectively. In the *har1-3* allele an 18-nucleotide deletion removes 6 amino acids from kinase domain XI, which presumably affects its activity.

Database analysis revealed a close similarity of the *HAR1* protein to the *Arabidopsis* protein *CLAVATA1* and to proteins predicted by conceptual translation of two soybean cDNAs²¹. Overall, the *HAR1* protein shares a 48.9%, 70.4% and 71.1% identity to *Arabidopsis* *CLV1* and the predicted soybean proteins *Glycine max* *CLV1A* and *CLV1B*, whereas local similarity in the C-terminal kinase domains is 74.8% to *CLV1*, 89.1% to *G. max* *CLV1A* and 91.5% to *G. max* *CLV1B*. The *G. max* *CLV1B* gene may be the soybean orthologue of *HAR1*.

In pea, two loci, *P. sativum* *SYM28* and *SYM29*, mediate shoot regulation of nodule number²², and pea mutants exhibit similar root and nodulation phenotypes as *L. japonicus* *har1* (refs 11, 22). However, *P. sativum* *sym28* also causes stem fasciation²², a phenotype not observed in *L. japonicus* *har1* and *P. sativum* *sym29* mutant lines^{11,22}. To test the hypothesis that *P. sativum* *SYM29* is the orthologue of *L. japonicus* *HAR1*, the pea homologue of *HAR1* was cloned from the parental wild-type cultivar 'Frisson' using

a polymerase chain reaction (PCR) approach. There was co-segregation between the *P. sativum* *SYM29* locus and the cloned *HAR1* homologue, and sequencing of the *P. sativum* *SYM29* gene from 13 independent *P. sativum* *sym29* mutant lines identified 9 mutant alleles of the gene carrying either nonsense or missense changes (Fig. 3 and Table 1; see also Supplementary Information).

Pisum sativum *SYM29* was found to encode a putative serine/threonine receptor-like kinase (*SYM29*) with a predicted relative molecular mass of 107.8K. Comparison of the conceptual *HAR1* and *P. sativum* *SYM29* proteins showed an overall amino acid identity of 77% (Supplementary Information). The premature stop codons identified in pea mutants P89/P94/P117, P116 and P122 are predicted to generate truncated *SYM29* proteins of 909, 666 and 341 amino acids, respectively, which lack part of, or the entire kinase domain (Fig. 3 and Table 1). In mutants P90 and P91, substitutions of the invariant glycine residues, GKGGAGIVY to GKRGAGIVY and GKGGAEIVY, in the glycine-rich ATP-binding site of kinase domain I, respectively, are predicted to affect the catalytic activity of the *SYM29* protein. The kinase activation segment of domain VII is mutated in P87, substituting the invariant glycine residue DFGLAK to DFRLAK. Two independent mutations were found to be localized in the ninth and tenth repeat of the conserved LRR domain: in mutant P106, the predicted protein sequence MKSLMSLDL is changed to MKSLMSLN_L, whereas in the P88, P93 and P119 mutants, a C to T transition leads to substitution of one of the conserved leucine residues (MKSLMSLDL to MKSEMSLDL). Finally, mutant P118 carries a missense substitution, CDQDNRVITL to CDQDNRMITL, located between the paired cysteines and the first LRR repeat.

On the basis of strong phenotypic similarities between the mutant plants, co-segregation data, sequence conservation, and characterization of nine independent *P. sativum* *sym29* mutant alleles, we conclude that *HAR1* and *P. sativum* *SYM29* are orthologous genes. As *HAR1* and *SYM29* are potential components of a long distance signalling mechanism(s), we investigated their expression in various organs of *L. japonicus* and pea. Northern hybridization and quantitative PCR with reverse transcription (RT-PCR) demonstrated that expression of these genes is widespread. Transcripts were detected in roots, nodules, cotyledons, stems and leaves of *L. japonicus*, and in roots, nodules and leaves of pea plants (Fig. 4a, d). The steady-state level of *L. japonicus* *HAR1* transcript was not detectably altered by the presence or absence of either

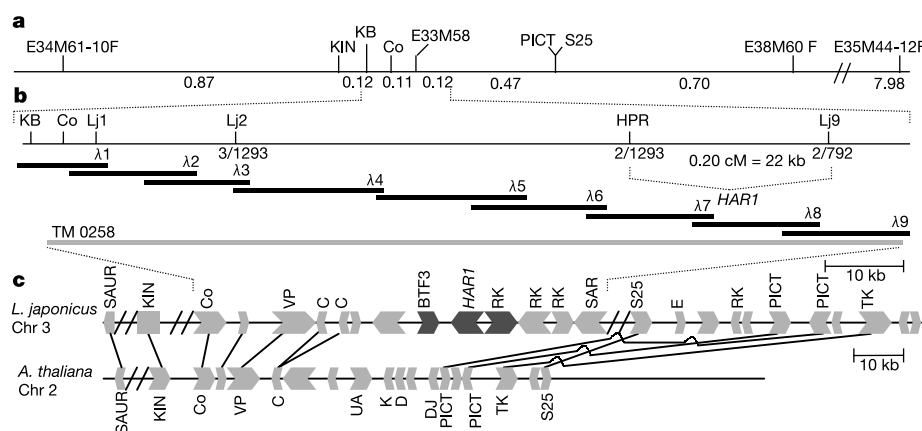


Figure 2 Map-based cloning of *HAR1*. **a**, Genetic map of the *HAR1* region with positions of linked AFLP and PCR-based markers above the line, and distances in cM below. **b**, Physical map of the λ (black bars) and TAC (grey bar) clones that were sequenced. Sequence-derived markers used to fine-map the *HAR1* locus and recombination events found with Lj2, HPR and Lj9 are indicated. **c**, Syntenic regions between *L. japonicus* chromosome 3 and a duplicated region on *Arabidopsis* chromosome 2 and 4 flank the *HAR1* locus located within a non-syntenic cluster of four receptor kinase genes (RK).

Candidate genes located in the 22-kb region delimited by recombinations are highlighted in grey. Lines connect homologous genes. KIN, kinesin; Co, coatomer alpha subunit; VP, vacuolar proton ATPase; C, papain-like cysteine proteinase; BTF3, putative transcription factor BTF3; SAR, SAR DNA-binding protein; S25, 40S ribosomal protein S25; E, fatty acid elongase; PICT, phosphatidylinositol/phosphatidylcholine transfer protein; TK, dihydrofolate reductase-thymidylate synthase; UA, ubiquitin-activating enzyme; K, protein kinase; D, dehydrin; DJ, DNAJ protein; unlabelled, hypothetical protein.

symbiotic bacteria or KNO₃, indicating that the *HAR1* gene is unlikely to be regulated at the transcriptional level by either a symbiosis-derived signal or the nitrogen status of the plant (Fig. 4a–c).

We describe a role of HAR1 as a component of the autoregulatory feedback mechanism required to maintain homeostasis of non-symbiotic and symbiotic root development in legumes. HAR1 receptor kinase orchestrates root development by mediating a long-range shoot to root signalling mechanism(s), and thus seems to be a vital part of a regulatory circuit that integrates root development with events in the plant shoot. The cellular changes observed in *har1* mutant roots¹¹ suggest that cells in the root meristematic zone and/or the pericycle cells are targets for the control system. Together, these observations indicate that the original role of *HAR1*; that is, the orchestration of root growth, has been recruited during evolution to serve a function during root nodule development. Although post-transcriptional control of HAR1 activity cannot be excluded, constitutive expression of the *HAR1* gene in *L. japonicus* indicates that HAR1 is unlikely to be a limiting factor in this mechanism; a role probably carried out by a ligand or interacting protein(s) triggering HAR1 receptor signalling. On the basis of split root and grafting experiments in hypernodulating soybean and pea, a model for autoregulation has

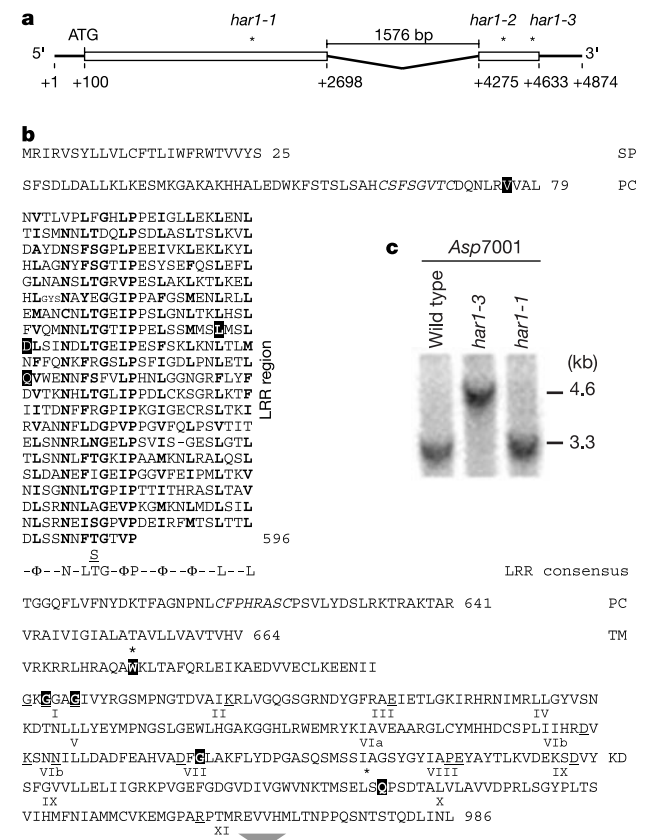


Figure 3 Structure of the *HAR1* gene and domains of the HAR1 protein. **a**, Intron-exon structure and transcription initiation point of *HAR1*. Asterisks mark nucleotide changes in the *har1-1*, *har1-2* and *har1-3* alleles. **b**, The amino acid sequence of HAR1 arranged in protein domains. Italic font indicates paired cysteines (PC); bold font indicates conserved LRR residues appearing at $\geq 50\%$ frequency; Φ , hydrophobic residues; underlined letters indicate residues conserved in protein kinase domains (KD)²⁹. TM, transmembrane; SP, signal peptide. Asterisks indicate stop codons in the *har1-1* and *har1-2* sequence. The grey triangle indicates a deletion in *har1-3*; black boxes indicate mutations in *P. sativum* *SYM29* alleles. **c**, Southern hybridization with a *HAR1*-specific probe visualizing an RFLP resulting from deletion of the Asp700L site in the *har1-3* allele. Only one hybridization signal was obtained with the 5' specific probe.

Table 1 Summary of <i>har1</i> and <i>sym29</i> mutant alleles			
Signature	Allele	Mutation	Species
1	<i>har1-1</i>	TGG $\rightarrow$ TAG; W676 $\rightarrow$ stop	Lj
2	<i>har1-2</i>	AAG $\rightarrow$ TAG; Q919 $\rightarrow$ stop	Lj
3	<i>har1-3</i>	REVVMH 964–969 deletion	Lj
a	P118	GTG $\rightarrow$ ATG; V72M	Pea
b	P88, P93, P119	CTC $\rightarrow$ TTC; L290F	Pea
c	P106	GAT $\rightarrow$ AAT; D294N	Pea
d	P122	CAG $\rightarrow$ TAG; Q342 $\rightarrow$ stop	Pea
e	P116	TGG $\rightarrow$ TAG; W667 $\rightarrow$ stop	Pea
f	P90	GGA $\rightarrow$ AGA; G695R	Pea
g	P91	GGG $\rightarrow$ GAG; G698E	Pea
h	P87	GGA $\rightarrow$ AGA; G831R	Pea
i	P89, P94, P117	CAG $\rightarrow$ TAG; Q910 $\rightarrow$ stop	Pea

Mutations, predicted protein truncations and amino acid substitutions of three *har1* and nine *P. sativum* *sym29* mutant alleles are shown. The position of changes in the proteins is shown in Fig. 3 and Supplementary Information as signatures 1–3 and a–i. Lj, *Lotus japonicus* Gifu; Pea, *Pisum sativum* Frisson.

been proposed where the shoot monitors nodule number through perception of a root-derived signal. As part of a negative feedback mechanism, shoot-derived signal(s) would limit further nodulation events^{3,4,8}. Our grafting and expression studies in *L. japonicus* show that HAR1 is required for perception of the proposed root signal and that interaction of HAR1 with a shoot-localized ligand or other associating proteins²³ is required for transduction of the root signal into an autoregulatory response inhibiting lateral organ development on the root. This would explain why root-expressed HAR1 is unable to compensate for the *har1* mutation in grafted shoots. The graft experiments also exclude a role for HAR1 in production of the root-derived signal and for perception of the shoot-derived negative signal in the root. Future characterization of the pea *NOD3* and *SYM28* genes, shown by grafting to mediate root and shoot control of hypernodulation, respectively, could identify these components. In *Arabidopsis*, CLV1 is required for cell differentiation at the flanks of the shoot apical meristem, thus maintaining homeostasis of the stem cell population and organ initiation in the shoot²⁴. HAR1 apparently parallels this function in the root by balancing cell

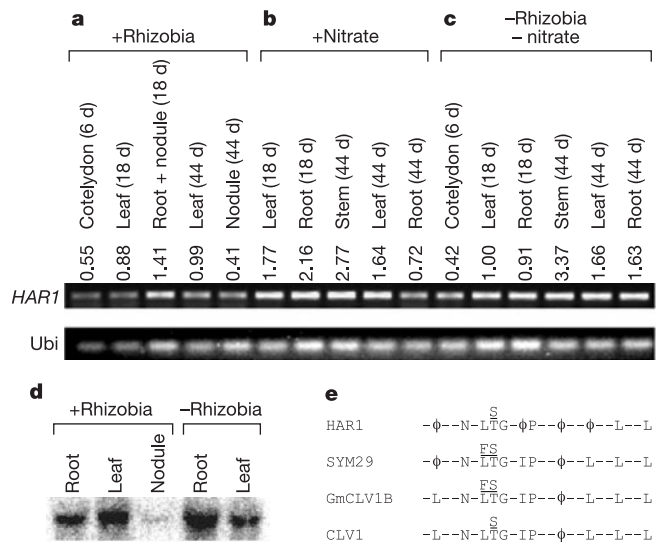


Figure 4 Expression of the *HAR1* and *P. sativum* *SYM29* genes in different plant organs under symbiotic and non-symbiotic conditions. **a**, In the presence of rhizobia; **b**, in the absence of rhizobia and the presence of 1 mM KNO₃; **c**, in the absence of rhizobia and KNO₃. The identity of the amplified sequences was confirmed by sequencing. Ubiquitin was used as an internal control and normalized values are indicated above individual lanes. **d**, Northern analysis showing *P. sativum* *SYM29* expression in symbiotically and non-symbiotically grown pea plants. **e**, Alignment of the consensus LRRs from HAR1, SYM29, *G. max* CLV1B and CLV1 proteins. Alternative amino acids in the same position are above and below the line (Φ , hydrophobic residues).

proliferation and differentiation in the root apical meristem and by regulating the extent of lateral root and nodule formation¹¹. A direct functional overlap between pathways controlling shoot and root cell differentiation is not supported by the *har1*, *P. sativum sym29* and *clv1* mutant phenotypes. An abnormal shoot phenotype was not detected in *har1* and *P. sativum sym29* mutants, and an abnormal root phenotype was not reported for *Arabidopsis clv1* mutants. *Arabidopsis CLV1* belongs to a large gene family of 150 members²⁴, and it is likely that other members of the *HAR1* and *P. sativum SYM29* gene families encode the *L. japonicus* and pea counterparts of the *A. thaliana CLV1* gene. Isolation of two related cDNAs, *G. max CLV1B* and *CLV1A*, from soybean²¹ supports this view. The micro-synteny (Fig. 2) between *L. japonicus* and *Arabidopsis* does not provide a simple explanation for the molecular evolution of the *HAR1* locus, and at present a *CLV1* gene has not been identified in legumes. Cross-talk between the *HAR1* and *CLV1* pathways is indicated by a mutation in the *SYM28* hypernodulation locus that causes stem fasciation²², a phenotype observed in *Arabidopsis clv1* mutants¹⁸. This observation and the overall conservation of LRR consensus sequences for *HAR1*, *P. sativum SYM29*, *G. max CLV1B* and *CLV1* (Fig. 4e) raises the possibility that small protein molecule(s) participate in long-distance signalling in plants. The 12 allele-specific modifications defining functional motifs in the extracellular receiver domain or the cytoplasmic kinase domains of the *HAR1* and *P. sativum SYM29* serine/threonine receptor kinases will be important for future investigation of this working hypothesis. □

Methods

Plant material

The *har1-1*, *har1-2* and *har1-3* (*sym16*) mutants were identified in two independent screens for symbiotic mutants^{9,10}. Plants were grown with or without *Mesorhizobium loti* NZP2235 as previously outlined²⁵. The pea *P. sativum sym29* complementation group was described previously²², and additional mutants have since been ascribed to the complementation group. The pea line JI281 was obtained from JIC Norwich-UK (<http://www.jic.bbsrc.ac.uk/germplas/pisum/index.htm>).

Grafting

Plants were sterilized and germinated as previously described²⁵. Seedlings were transferred to plates containing 1/4 concentration of agar solidified B&D nutrients²⁵ and grown for 4–6 days at 21 °C in a 16/8-h light/dark regime, in the presence of either 0.1 mM or 1 mM KNO₃, under symbiotic or non-symbiotic conditions, respectively. Shoot–root junctions were grafted in a plastic tube (diameter of 0.6 mm) and fixed to 1/4 B&D plates with agar. After 14 days of recovery plants were transferred to Magenta jars containing Leca tile²⁵, and were watered with 50 ml of 1/4 B&D. When inoculated, we applied 5 ml of a 2-day-old *M. loti* NZP2235 culture diluted 1:1,000. Plants were grown as described above and analysed after 1 to 2 months. Under symbiotic conditions, between 9 and 15 individual grafts were analysed for each shoot–root combination.

Mapping population

Intraspecific and interspecific F₂ mapping populations were established by crossing a *L. japonicus* Gifu *har1-3* mutant to wild-type *L. japonicus* ecotype MG20 (ref. 26) and wild-type *L. filicaulis*¹⁵. F₂ plants homozygous for the *har1-3* mutant allele were identified after screening for the hypernodulation mutant phenotype. A total of 792 homozygous F₂ mutant plants were analysed in the *har1-3* × *L. filicaulis* population and 501 homozygous F₂ mutant plants in the *har1-3* × MG20 mapping population.

Positional cloning

AFLP and bulked segregant analysis were performed using the *EcoRI*–*MseI* restriction enzyme combination^{27,28}. Initially, the closely linked AFLP marker E33M58 was identified and end-sequencing of a corresponding λ clone identified a H⁺ ATPase gene (Fig. 2). Together with sequence information from a *SAUR* gene cluster identified in the *Lotus* bacterial artificial chromosome (BAC) clone corresponding to the E34M61–10F marker, this indicated microsynteny to a region on chromosome 2 of *Arabidopsis* between positions 9,250,000 and 9,430,000 (BACs F7024, F3K23, F2G1). Taking advantage of this microsynteny and gene sequence conservation, four new PCR markers (KIN, Co, PICT and S25; see legend to Fig. 2) were found. Two of these markers, PICT and S25, mapped to the opposite side of *HAR1* as E34M61–10F, KIN, Co and E33M58. Using these flanking markers we assembled a contig of TAC and λ clones. Finally, a TAC clone (MG20) and λ contig (Gifu) covering 105 kb between the closest linked flanking markers were sequenced. Single nucleotide polymorphisms (SNPs) and microsatellites (Lj2, HPR, Lj9) derived from the ecotype-specific DNA sequences were then used to fine map *HAR1* to a 0.2-cM region (22 kb) delimited by recombination end points (Fig. 2b). The λ contig was obtained by screening a *L. japonicus* Gifu λ Fix II library, and the TAC clone sequence was provided by the Kazusa DNA Research Institute.

cDNA isolation

One partial cDNA clone of 3,034 bp was identified by screening a *L. japonicus* Gifu leaf cDNA library with a *HAR1* gene-specific probe. The 265-bp 5' fragment was amplified using the SMART RACE cDNA Amplification Kit (Clontech) and a full-length cDNA clone was assembled. The Southern probe is located between positions 1224 to 2845.

Cloning of SYM29 and co-segregation analysis

Primers for PCR amplification of a *P. sativum SYM29* gene fragment were designed on the basis of comparison of the *L. japonicus HAR1* sequence with the soybean homologues AF197946 and AF197947. The approximately 530 base pairs PCR product of the forward primer 5'-CCTCAACATCTCYCACACSTCTCTC-3' and the reverse primer 5'-CGTTGAWGGAGAGRTCCAGTGACATG-3' (annealing temperature 55 °C) was cloned, sequenced and used as a probe to isolate a genomic clone. On the basis of the sequence of this clone, the sequences of the gene in Frisson and JI281 were determined by a PCR amplification/sequencing approach, and SNP-based primers were designed for co-segregation mapping. Co-segregation was found in 26 F₂ *P. sativum sym29* mutant plants of a mapping population, whereas independent segregation was found in *sym28* mutants. We determined the sequences of the *sym29* mutant alleles by sequencing PCR products covering the gene.

Expression analysis

Different tissues were collected at 18 and 44 days after infection, and the level of *HAR1* steady-state messenger RNA was determined by real-time PCR and northern hybridization. Total RNA was extracted by centrifugation in a caesium chloride cushion. The RNA was treated with DNase I (Boehringer) to degrade contaminating genomic DNA. The DNase I enzyme was removed by phenol:chloroform extraction and the RNA was precipitated and re-suspended in 10 µl DEPC H₂O. First strand cDNA was prepared using Superscript II RNase H⁻ reverse transcriptase from Invitrogen. Quantitative real-time PCR was performed in a PE GeneAmp 5700 SDS with the quantitative PCR Master Mix for SYBR Green I Kit from Eurogentec. Correction for differences in template amounts was done using ubiquitin as a reference. The *HAR1* transcript levels indicated above individual lanes were obtained using the relative quantification method provided with the GeneAmp 5700 sequence detection system manual. The value for leaf at 18 d without rhizobia or nitrogen was used as a calibrator.

Computer analysis

Sequences were analysed by BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) and PredictProtein (<http://www.embl-heidelberg.de/predictprotein/predictprotein.html>). The signal peptide was predicted by SignalP and the transmembrane region was predicted by TMHMM at <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>.

Received 19 July; accepted 11 October 2002; doi:10.1038/nature01207.

Published online 6 November 2002.

- Spaink, H. P. Root nodulation and infection factors produced by rhizobial bacteria. *Annu. Rev. Microbiol.* **54**, 257–288 (2000).
- Nutman, P. S. Studies on the physiology of nodule formation. III. Experiments on the excision of root tip and nodules. *Ann. Bot.* **16**, 81–102 (1952).
- Pierce, M. & Bauer, W. D. A rapid regulatory response governing nodulation in soybean. *Plant Physiol.* **73**, 286–290 (1983).
- Kossak, R. M. & Bohlool, B. B. Suppression of nodule development of one side of a split root system of soybeans caused by prior inoculation of the other side. *Plant Physiol.* **75**, 125–130 (1984).
- Carroll, B. J. & Mathews, A. *Molecular Biology of Symbiotic Nitrogen Fixation* (ed. Gresshoff, P.) 159–180 (CRS, Boca Raton, 1990).
- Vasse, J., de Billy, F. & Truchet, G. Abortion of infection during the *Rhizobium meliloti*-alfalfa symbiotic interaction is accompanied by a hypersensitive reaction. *Plant J.* **4**, 555–566 (1993).
- Pennetsa, R. V. & Cook, D. R. A legume ethylene-insensitive mutant hyperinfected by its *Rhizobial* symbiont. *Science* **275**, 527–530 (1997).
- Caetano-Anolles, G. & Gresshoff, P. M. Plant genetic control of nodulation. *Annu. Rev. Microbiol.* **45**, 345–382 (1991).
- Schauser, L. et al. Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of *Lotus japonicus*. *Mol. Gen. Genet.* **259**, 414–423 (1998).
- Szczygłowski, K. et al. Nodule organogenesis and symbiotic mutants of the model legume *Lotus japonicus*. *Mol. Plant-Microbe Interact.* **11**, 684–697 (1998).
- Wopereis, J. et al. Short root mutant of *Lotus japonicus* with a dramatically altered symbiotic phenotype. *Plant J.* **23**, 97–114 (2000).
- Carroll, B. J., McNeil, D. L. & Gresshoff, P. M. Isolation and properties of soybean [*Glycine max* (L.) Merr.] mutants that nodulate in the presence of high nitrate concentrations. *Proc. Natl Acad. Sci. USA* **82**, 4162–4166 (1985).
- Delves, A. C. et al. Regulation of the soybean–*Rhizobium* nodule symbiosis by shoot and root factors. *Plant Physiol.* **82**, 588–590 (1986).
- Sheng, C. & Harper, J. E. Shoot versus root signal involvement in nodulation and vegetative growth in wild-type and hypernodulating soybean genotypes. *Plant Physiol.* **113**, 825–831 (1997).
- Sandal, N. et al. A genetic linkage map of the model legume *Lotus japonicus* and strategies for fast mapping of new loci. *Genetics* **161**, 1673–1683 (2002).
- Sato, S. et al. Structural analysis of the *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Res.* **8**, 311–318 (2001).
- Nakamura, Y. et al. Structural analysis of a *Lotus japonicus* genome. II. Sequence features and mapping of sixty-five TAC clones which cover the 6.5-Mb regions of the genome. *DNA Res.* **9**, 63–70 (2002).
- Clark, S. E., Williams, R. W. & Meyerowitz, E. M. The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in *Arabidopsis*. *Cell* **89**, 575–585 (1997).
- Shiu, S.-H. & Blecker, A. B. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proc. Natl Acad. Sci. USA* **11**, 10763–10768 (2001).
- Trotochaud, A. E., Hao, T., Wu, G., Yang, Z. & Clark, S. E. The *CLAVATA1* receptor-like kinase requires

- CLAVATA3 for its assembly into a signalling complex that includes KAPP and Rho-related protein. *Plant Cell* **11**, 393–405 (1999).
21. Yamamoto, E., Karakaya, H. C. & Knap, H. T. Molecular characterization of two soybean homologs of *Arabidopsis thaliana* CLAVATA1 from the wild type and fasciation mutant. *Biochem. Biophys. Acta* **1491**, 333–340 (2000).
 22. Sagan, M. & Duc, G. *Sym28* and *Sym29*, two new genes involved in regulation of nodulation in pea (*Pisum sativum* L.). *Symbiosis* **20**, 229–245 (1996).
 23. Torii, K. U. Receptor kinase activation and signal transduction in plants: an emerging picture. *Curr. Opin. Plant Biol.* **3**, 361–367 (2000).
 24. Clark, S. E. Cell signalling at the shoot meristem. *Nature Rev.* **2**, 276–284 (2001).
 25. Handberg, K. & Stougaard, J. *Lotus japonicus*, an autogamous diploid legume species for classical and molecular genetics. *Plant J.* **2**, 487–496 (1992).
 26. Kawaguchi, M. *Lotus japonicus* 'Miyakojima' MG20: An early-flowering accession suitable for indoor handling. *J. Plant Res.* **113**, 507–509 (2000).
 27. Vos, P. et al. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**, 4407–4414 (1995).
 28. Michelmore, R. W., Paran, I. & Kesseli, R. V. Identification of markers linked to disease-resistance genes by bulked segregant analysis. A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA* **88**, 9828–9832 (1991).
 29. Hanks, S. K. & Quinn, A. M. Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **200**, 38–62 (1988).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements This work was funded by the Danish Ministry of Food, Agriculture and Fishery. Support was also provided by the National Science Foundation and the Kazusa DNA Research Institute Foundation. The Danish Agricultural and Veterinary Research Council supports L.K. We thank K. Keegstra for administrative support during the course of the work, A. Albin for administrative assistance, and H. de Laremburgue for technical assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S. (e-mail: stougaard@mbio.au.dk). The sequences for the *L. japonicus* *HAR1* gene and mRNA are deposited in GenBank under accession numbers AJ495843 and AJ495844, respectively; the *P. sativum* *SYM29* gene is deposited under accession number AJ495759. The sequences for TAC LT09E01/TM0258 are deposited in GenBank under accession numbers AP005665, AP005666 and AP005667.

HAR1 mediates systemic regulation of symbiotic organ development

Rieko Nishimura*†‡, Masaki Hayashi†§, Guo-Jiang Wu†||¶, Hiroshi Kouchi||, Haruko Imaizumi-Anraku||, Yasuhiro Murakami||, Shinji Kawasaki||, Shoichiro Akao||, Masayuki Ohmori*, Mamoru Nagasawa§, Kyuya Harada§ & Masayoshi Kawaguchi¶#

* Department of Life Science, Graduate School of Arts and Sciences, The University of Tokyo, Komaba, Meguro-ku, Tokyo 163-8902, Japan

§ Faculty of Horticulture, Chiba University, 648 Matsudo, Matsudo, Chiba 271-8510, Japan

|| National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

¶ Department of Environmental Sciences, Faculty of Science, Niigata University, Ikarashi, Niigata City 950-2181, Japan

Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation, 4-1-8 Honcho, Kawaguchi City, Saitama 332-0012, Japan

† These authors contributed equally to this work

Symbiotic root nodules are beneficial to leguminous host plants; however, excessive nodulation damages the host because it interferes with the distribution of nutrients in the plant. To keep a steady balance, the nodulation programme is regulated systemically in leguminous hosts^{1,2}. Leguminous mutants that have lost this ability display a hypernodulating phenotype.

Through the use of reciprocal and self-grafting studies using *Lotus japonicus* hypernodulating mutants, *har1* (also known as *sym78*)³, we show that the shoot genotype is responsible for the negative regulation of nodule development. A map-based cloning strategy revealed that *HAR1* encodes a protein with a relative molecular mass of 108,000, which contains 21 leucine-rich repeats, a single transmembrane domain and serine/threonine kinase domains. The *har1* mutant phenotype was rescued by transfection of the *HAR1* gene. In a comparison of *Arabidopsis* receptor-like kinases, *HAR1* showed the highest level of similarity with *CLAVATA1* (*CLV1*)⁴. *CLV1* negatively regulates formation of the shoot and floral meristems through cell–cell communication involving the *CLV3* peptide⁵. Identification of hypernodulation genes thus indicates that genes in leguminous plants bearing a close resemblance to *CLV1* regulate nodule development systemically, by means of organ–organ communication.

Symbiosis in legume can exist as a delicate balance between hosts and microbes. To prevent excessive nodulation, leguminous plants have negative regulatory systems; for example, the ethylene signalling system negatively regulates rhizobial infection⁶. Similarly, there exists a systemic feedback regulatory system controlling nodule development. This process, in which developed nodules or nodule primordia suppress the emergence of further nodules⁷, is termed autoregulation^{1,2,8}. This feedback system greatly restricts the zone of nodulation in the root. Furthermore, application of Nod factors (lipochitin oligosaccharides) systemically elicits feedback regulation in the *Vicia* split root system⁹. Control of nodulation by systemic feedback is of great interest with respect to the flexibility of plant organogenesis; however, little is known about the molecular mechanisms underlying this process.

We have described previously two ethylmethane sulphonate (EMS)-induced *L. japonicus* mutants with an increased number of nodules and wider nodulation zones: *L. japonicus sym77* (*astray*)¹⁰ and *sym78* (ref. 3). The latter mutant is allelic to hypernodulating *har1* mutants^{11–13}, and shows an enhanced mycorrhizal phenotype¹⁴. Here we report impaired systemic regulation by *har1* mutants and molecular identification of the *HAR1* gene.

We used grafts involving Gifu B-129 (wild type) and the *har1-5* mutant line to determine the role of shoot and root factors in the hypernodulating phenotype. Grafting a *har1-5* shoot onto a wild-type root resulted in increased nodule numbers and an extended nodulation zone (Table 1). On the other hand, grafting a wild-type shoot onto a *har1-5* root suppressed the hypernodulating phenotype and restricted the nodulation zone (Table 1). These results indicate that the *har1* mutant, similar to soybean *nts1* and pea hypernodulating mutants^{15–18}, is unable to produce an autoregulation signal from the shoots. We then evaluated the systemic regulation of *har1* mutants using split-root experiments (see Supplementary Fig. 1). Lateral roots (facilitated by cutting off the main root tip) were split equally. Root A was first inoculated with *Mesorhizobium loti* MAFF (Ministry of Agriculture, Forestry and Fisheries, Japan) 303099 at a concentration of 10^7 cells ml⁻¹, followed by inoculation of root B 8 days later. Five weeks after the second inoculation, the number of nodules on root B was only 29% of that on root A in wild-type plants, indicating systemic nodule

Table 1 Grafting experiments with *har1* mutants

Shoot/root genotype	<i>n</i>	Nodule number	Nodulation zone ratio
Wild type/wild type	10	10.70 ± 3.16	0.14 ± 0.04
Wild type/ <i>har1-5</i>	14	12.29 ± 3.27	0.16 ± 0.08
<i>har1-5</i> /wild type	10	46.90 ± 13.65	0.65 ± 0.14
<i>har1-5</i> / <i>har1-5</i>	8	43.75 ± 15.02	0.63 ± 0.10

Nodules and nodule primordia of more than 0.4 mm in diameter on grafted seedlings were scored one month after inoculation. The nodulation zone is based on the distance between the nodules developed at the uppermost and lowermost position on the main root. The ratio of the nodulation zone is calculated by dividing the nodulation zone length by the main root length. Values are given with the indicated standard deviation. *n*, number of plants tested.

‡ Present address: Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720-3206, USA

suppression on root B owing to the preceding nodulation on root A. In contrast, the numbers of nodules on roots A and B were almost the same in *har1* mutants (see Supplementary Fig. 1). These results are similar to the outcome of tests on the soybean *ntsl* mutant¹⁹, and suggest that *har1* mutants are deficient in systemic suppression of nodule development.

The *har1* mutation was mapped to the long arm of chromosome 3. Assuming microsynteny with the soybean *ntsl* genomic region, we searched for *L. japonicus* orthologues of soybean restriction-fragment length polymorphism (RFLP) markers, expected to be located near the *NTS1* gene, in a database of *L. japonicus* expressed sequence tags (ESTs), and used them for linkage analysis. Consequently we found that *L. japonicus* *SUT1*, an orthologue of *Glycine max* 221 (sugar transporter)²⁰, is tightly linked to the *har1* gene at a distance of 0.8 cM (Fig. 1a). A transformation-competent artificial chromosome (TAC) clone, *L. japonicus* T25N16, carrying the *L. japonicus* *SUT1* gene, also contains the gene for fructose-bisphosphate aldolase; the soybean orthologue *G. max* 036 (ref. 20) has been mapped to the vicinity of *ntsl* (P. M. Gresshoff, personal communication). An amplified-fragment length polymorphism (AFLP) co-dominant marker, E2M16-C160, found by bulked segregant analysis was mapped near *har1* at a distance of 0.11 cM (Fig. 1a). Starting with the E2M16-C160 marker, we isolated three bacterial artificial chromosome (BAC) clones and sequenced a genomic region towards the *har1* locus.

By sequencing the BAC clone 259.12D and the DNA polymorphism between Gifu and Miyakojima, we found that the *har1* locus lies between E2M16-C160 and an indel marker, G3INS. This 72-kilobase (kb) genomic region (Fig. 1b) includes ten possible genes, one of which was altered in two *har1* alleles that we sequenced. We used a 12.5-kb genomic fragment containing this open reading frame (ORF), together with 3.7 kb of upstream and 4.0 kb of downstream sequence, in complementation tests by transforming it to the *har1* mutants. Using *Agrobacterium tumefaciens*-mediated hypocotyl transformation we obtained 10 and 25 independent transformants for *har1-4* for *har1-5*, respectively. Among them, all lines we examined exhibited nodule numbers and nodulation zone ratios that were identical to those of wild-type Gifu plants (Fig. 2; see also Supplementary Table 1). Southern hybridization indicated that these transgenic plants harbour either one or two transgenes (data not shown). Shoot and root growth of the transformants also recovered to levels approaching those in wild-type Gifu plants, although recovery was not complete in some of the lines. As nodule number and nodulation zone ratio are the two parameters of central

importance for the hypernodulating phenotype, we conclude that transfection of an intact *HAR1* gene can completely rescue the hypernodulating phenotype of *har1* mutants.

The *HAR1* gene contains a long ORF that encodes a receptor-like kinase comprising 987 amino acids, containing leucine-rich repeats (LRRs) (Fig. 3). The relative molecular mass of the deduced protein is 108,000. The N terminus of the *HAR1* gene product has a potential signal peptide. The putative extracellular domain flanked by paired cysteine residues consists of 21 tandem LRRs and 2 partial repeats (see Supplementary Fig. 2). Following the single transmembrane domain, the intracellular region shows a high level of similarity with serine/threonine protein kinases. The severe *har1-4* allele³ was due to an amino acid substitution (leucine (CTC) to phenylalanine (TTC)) at the conserved position that connects the β -sheet and loop. This indicates that the severe phenotype of this allele may be due to reduced ligand binding. The missense mutation of the *har1-5* allele (glutamic acid (GAG) to lysine (AAG)) was found in region III of the serine/threonine kinase domain. Sequencing of a soybean orthologue, *G. max* *CLV1B*²¹, showed that the EMS-induced hypernodulating mutant En6500 (ref. 22), which

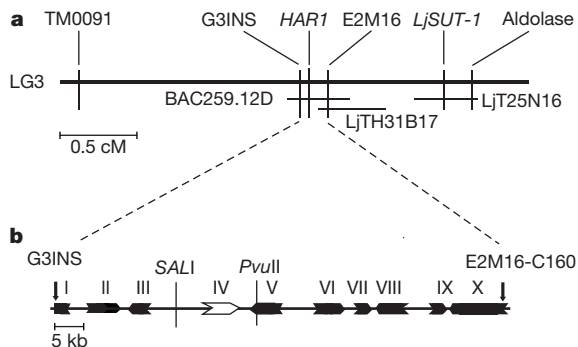


Figure 1 Map-based isolation of the *L. japonicus* *HAR1* gene. **a**, A genetic linkage map of the region of *HAR1* in linkage group 3. **b**, Physical map of the 72.5-kb region sequenced within BAC259.12D. Arrows indicate the positions of the E2M16 and G3INS markers. We carried out gene modelling using GENSCAN. Homology search of the predicted genes was performed by BLASTX. I, II and III, S-receptor kinase (T05754); IV, *HAR1*; V, putative transcription factor (AY084841); VI, hypothetical protein (C84809); VII, calmodulin-binding protein (2806318A); VIII, hypothetical protein (AC018727); IX, papain-like cysteine protease isoform II (AF138265); X, vacuolar proton ATPase subunit (H84600).

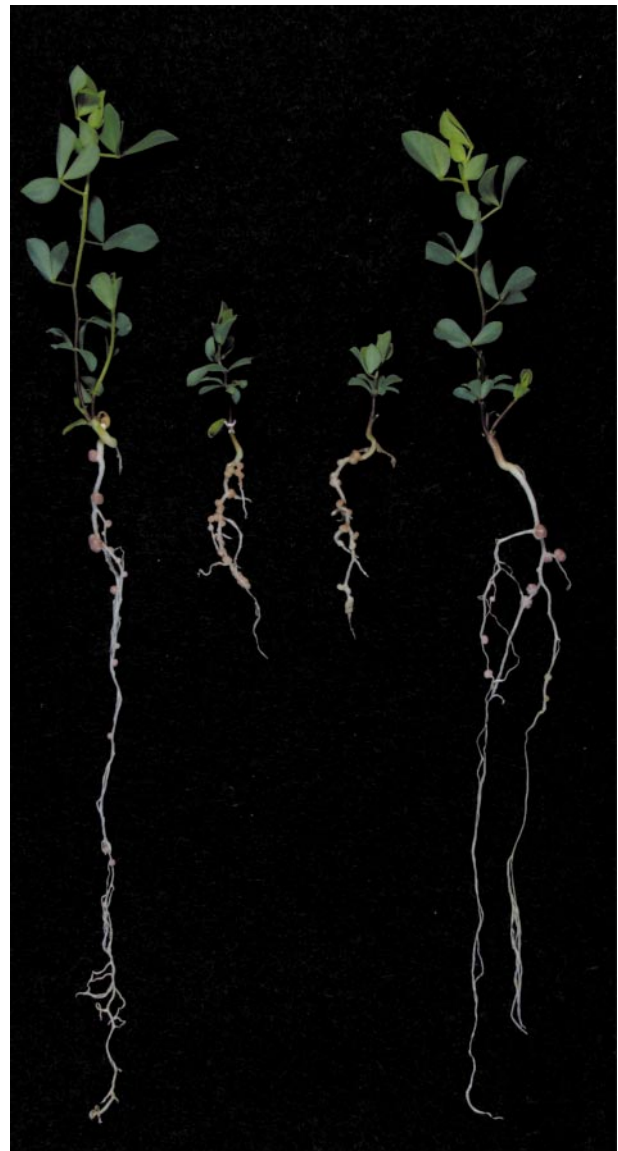


Figure 2 Complementation of a *har1* mutant with a 12.5-kb genomic fragment containing the complete *HAR1* gene. From left to right: wild-type Gifu, *har1-5*, *har1-5* transfected with an empty vector, and *har1-5* transfected with the 12.5-kb genomic fragment.

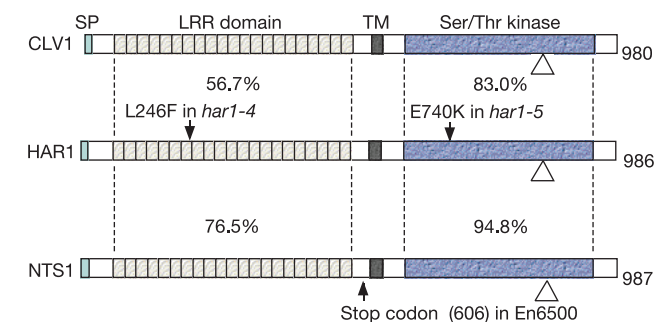


Figure 3 Schematic diagram and comparison of three LRR receptor kinases. SP, putative signal peptide; LRR domain, 21 tandem LRR domain; TM, single transmembrane domain; Ser/Thr kinase, serine/threonine kinase domain. Triangles indicate the locations of introns. Mutations are indicated in the HAR1 and NTS1 (G. max CLV1B, the soybean gene product showing the highest similarity with HAR1) diagrams. Percentage values represents amino acid identity.

is allelic to the *nts1* mutant, has a stop codon (lysine (AAA) to (TAA)) near the transmembrane domain (Fig. 3; see also Supplementary Fig. 2).

Out of 216 *Arabidopsis thaliana* receptor-like kinases that contain LRRs, the HAR1 gene product has the highest homology to CLV1 (ref. 4), which functions in the organization of shoot apical meristem. The amino acid sequence of HAR1 shares 56.7% and 83.0% identity with the LRR and kinase domains of CLV1, respectively. These are much higher levels of identity than those shown by the phytosulphokine²³ and systemin receptors²⁴, which are small peptide-binding receptors with LRR and kinase domains. A phylogenetic tree, based on the LRR domains that have pivotal roles in ligand binding, demonstrates the close relationship of leguminous HAR1 and NTS1 receptor-like kinases to CLV1 (see Supplementary Fig. 3).

The shoot apex was not altered in the *har1* mutants, despite the close structural relationship between HAR1 and CLV1. This suggests that a functional orthologue of CLV1 other than HAR1 is present in *L. japonicus*. Furthermore, this implies that the principal regulatory gene that organizes shoot and floral meristems was recruited to systemic nodule development during the course of evolution. We investigated HAR1 expression using northern hybridization. The 3.25-kb transcript was present in leaves, shoots, roots and flowers, and, at a much lower level, in nitrogen-fixing nodules (Fig. 4a). Notably, we did not detect expression in shoot apices, suggesting again that HAR1 differs from CLV1. Genomic Southern hybridization with the LRR region of the HAR1 gene suggested that HAR1 is present as a single copy in the genome of *L. japonicus* (Fig. 4b). Thus the CLV1 orthologue of *L. japonicus*, if present, may differ considerably in the primary structure compared with the HAR1 gene.

It has been proposed that autoregulation of nodule development consists of a root-derived signal and an autoregulation signal². The former is generated in nodules and nodule primordia, whereas the latter is produced in shoots, mainly leaflets¹⁷, on receiving the root-derived signal. Grafting experiments indicate that shoot genotype regulates nodule number (Table 1), but HAR1 transcripts are present in various organs including roots (Fig. 4a). Therefore, we cannot exclude completely the possibility that roots also participate in production of the autoregulation signal. A probable explanation for this is that most of the root-derived signal is immediately translocated to the shoots through the xylem stream, where the HAR1 gene product may recognize the signal and further promote the synthesis of the autoregulation signal (see Supplementary Fig. 4). As HAR1 encodes a receptor-like kinase with an LRR domain, this may bind directly the root-derived signal and hence transduce the

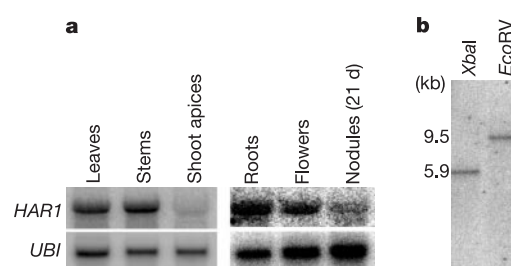


Figure 4 RNA and DNA blot analyses of HAR1 gene. **a**, Expressional analysis of HAR1 transcripts. Five micrograms of total RNA was prepared from various organs of *L. japonicus*, subjected to denaturing agarose gel electrophoresis, and blotted onto nylon membranes (Biodyne A). The membranes were hybridized with ³²P-labelled probes prepared from the LRR region of the HAR1 gene. Shoot apices were collected from 300 seedlings by cutting with a razor blade under magnification. UBI, polyubiquitin transcripts. **b**, Genomic DNA from Gifu B-129 was digested and blotted to a nylon membrane. The membrane was hybridized with the LRR region used in northern analysis and washed at high stringency.

signal by means of the intracellular kinase domain. The lack of systemic nodule suppression originating from formerly infected roots in the *har1* mutants (Supplementary Fig. 1) indicates a close relationship between HAR1 and the root-derived signal.

Molecular identification of the *L. japonicus* and soybean hypernodulation genes provides an indication as to the chemical nature of the presumptive root-derived signal. Comparison with another associated protein, CLV3, on forming a complex with CLV2 (ref. 5), may be informative. If the fact that the LRR domain of HAR1 is very similar to that of CLV1 is also taken into account, it seems probable that the HAR1 gene product binds a small peptide. One possibility is a CLV3-like peptide. Another possibility is peptides potentially encoded by *ENOD40* (refs 25, 26), the transcripts of which locate abundantly in the cells adjacent to xylem poles in roots at rhizobium-infected sites.

The hypernodulating gene is potentially agronomically important. A supernodulating line, En-bo-1-2, derived from En6500, produces a high yield owing to increased symbiotic nitrogen fixation through the later stages of growth (M. Takahashi, personal communication). Hence, the use of a series of alleles of the hypernodulating gene could make a significant contribution to molecular breeding in legumes. □

Methods

Plant materials

The mutants *har1-4* and *har1-5* (*L. japonicus sym78-1* and *sym78-2*) were isolated from the M2 generation of EMS-mutagenized seeds of *L. japonicus* Gifu B-129 as short-root mutants with an increased number of nodules. To generate a mapping population, the *har1-4* mutant was crossed to the accession Miyakojima MG-20 (ref. 27), and 448 lines of F₂ progeny were used for fine mapping.

Grafting experiments

Three-week-old seedlings were cut at the hypocotyl and the shoots were grafted to self or reciprocal roots using small plastic tubes. Grafted seedlings were grown in vermiculite immersed with B&D culture medium containing *M. loti* MAFF 303099. The seedlings were placed in moist containers and grown at 22 °C under a 16/8-h light/dark cycle.

Fine mapping

To identify the *har1* locus, AFLP markers from all six linkage groups were selected from the genetic linkage map of *L. japonicus*²⁸ at 5–10 cM intervals, and bulked segregant analyses with high-efficiency genome scanning (HEGS)²⁹ were performed. Primer sets showing differences between the bulks were applied to all of the individual plants of the F₂ progeny to determine their segregation and recombination values. The rest of the markers used were developed in this work. E2M16-C160 was amplified with 5'-GAATCAAGGGG ATCGATCTCAGTC-3' and 5'-TTAAGCTGGAAGAGATTATGCCAT-3'. G3INS was amplified with primers 5'-GTTGATAGCAAGACTTCAGC-3' and 5'-CCAAGCTTCT GTTGTGCATG-3'. Linkage analysis was performed on F₂ segregation data using the MAPMAKER/EXP version 3.0b program.

Cloning and sequencing *HAR1* and *NTS1*

A 72 kb region of BAC259.12D that encompasses the *HAR1* locus was sequenced in its entirety using a DNA Sequencing Kit (PE Applied Biosystems) with an automated DNA sequencer (ABI PRISM 3100; PE Applied Biosystems). *HAR1* complementary DNA was cloned from a cDNA library of *L. japonicus* shoots using DNA fragments of the first exon of the *HAR1* gene obtained by polymerase chain reaction (PCR) from BAC259.12D. After sequencing the cDNA clone, the full-length cDNA sequence was determined by 5' rapid amplification of cloned ends (RACE) on *L. japonicus* shoot messenger RNA. The soybean gene showing the highest similarity with *HAR1*, *G. max* *CLV1B*, was amplified from the genomic DNA of a soybean cultivar Enrei and of a hypernodulating mutant, En6500, by PCR with a specific primer set designed from the DNA Data Bank of Japan (DDBJ) database. They were sequenced using a DNA Sequencing Kit (PE Applied Biosystems) with an automated DNA sequencer.

Complementation experiments

A *SalI*–*PvuII* fragment (12.5 kb) containing the entire *HAR1* gene was excised from a BAC clone, 259.12D, and ligated into the *SalI*–*SmaI* site of binary vector pCambia1301 (Cambia). It was transformed into *A. tumefaciens* AGL1 (provided by R. A. Ludwig) and used for transformation of *har1-4* and *har1-5*. Transgenic plants were produced by the hypocotyl transformation procedure³⁰, with hygromycin B as a selective marker. Seeds from transgenic lines were germinated in vermiculite pots, inoculated with *M. loti* strain TONO, and examined for plant growth and nodulation behaviour 24 days after inoculation.

Received 22 July; accepted 21 October 2002; doi:10.1038/nature01231.
Published online 6 November 2002.

- Kosslak, R. M. & Bohlool, B. B. Suppression of nodule development of one side of a split-root system of soybeans caused by prior inoculation of the other side. *Plant Physiol.* **75**, 125–130 (1984).
- Gresshoff, P. M. & Caetano-Anollés, G. *Current Topics in Plant Molecular Biology: Plant Biotechnology and Development* (ed. Gresshoff, P. M.) 87–99 (CRC, Boca Raton, 1992).
- Kawaguchi, M. *et al.* Root, root hair, and symbiotic mutants of the model legume *Lotus japonicus*. *Mol. Plant-Microbe Interact.* **15**, 17–26 (2002).
- Clark, S. E., Williams, R. W. & Meyerowitz, E. M. The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in *Arabidopsis*. *Cell* **89**, 575–585 (1997).
- DeYoung, B. J. & Clark, S. E. Signaling through the *CLAVATA1* receptor complex. *Plant Mol. Biol.* **46**, 505–513 (2001).
- Penmetsa, R. V. & Cook, D. R. A legume ethylene-insensitive mutant hyperinfected by its rhizobial symbiont. *Science* **275**, 527–530 (1997).
- Nutman, P. S. Studies on the physiology of nodule formation. III. Experiments on the excision of root-tips and nodules. *Ann. Bot.* **16**, 79–101 (1952).
- Pierce, M. & Bauer, W. D. A rapid regulatory response governing nodulation in soybean. *Plant Physiol.* **73**, 286–290 (1983).
- van Brussel, A. A., Tak, T., Boot, K. J. & Kijne, J. W. Autoregulation of root nodule formation: signals of both symbiotic partners studied in a split-root system of *Vicia sativa* subsp. *nigra*. *Mol. Plant-Microbe Interact.* **15**, 341–349 (2002).
- Nishimura, R., Ohmori, M. & Kawaguchi, M. The novel symbiotic phenotype of enhanced-nodulating mutant of *Lotus japonicus*: *astray* is an early nodulating mutant with wider nodulation zone. *Plant Cell Physiol.* **43**, 853–859 (2002).
- Shauser, L. *et al.* Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of *Lotus japonicus*. *Mol. Gen. Genet.* **239**, 414–423 (1998).
- Szczygłowski, K. *et al.* Nodule organogenesis and symbiotic mutants of the model legume *Lotus japonicus*. *Mol. Plant-Microbe Interact.* **11**, 684–697 (1998).
- Wopereis, J. *et al.* Short root mutant of *Lotus japonicus* with a dramatically altered symbiotic phenotype. *Plant J.* **23**, 97–114 (2000).
- Solaiman, M. Z. *et al.* Characterization of mycorrhizas formed by *Glomus* sp. on root of hypernodulating mutants of *Lotus japonicus*. *J. Plant Res.* **113**, 443–448 (2000).
- Delves, A. C. *et al.* Regulation of the soybean–*Rhizobium* nodule symbiosis by shoot and root factors. *Plant Physiol.* **82**, 588–590 (1986).
- Hamaguchi, H., Kokubun, M. & Akao, S. Shoot control of nodulation is modified by the root in the supernodulating soybean mutant En6500 and its wild-type parent cultivar Enrei. *Soil. Sci. Plant Nutr.* **38**, 771–774 (1992).
- Sheng, C. & Harper, J. Shoot versus root signal involvement in nodulation and vegetative growth in wild-type and hypernodulating soybean genotypes. *Plant Physiol.* **113**, 825–831 (1997).
- Sagan, M. & Duc, G. *Sym28* and *Sym29*, two new genes involved in autoregulation of nodulation in pea (*Pisum sativum* L.). *Symbiosis* **20**, 229–245 (1996).
- Olsson, J. E., Nakao, P., Bohlool, B. B. & Gresshoff, P. M. Lack of systemic suppression of nodulation in split root systems of supernodulating soybean (*Glycine max* [L.] Merr.) mutants. *Plant Physiol.* **90**, 1347–1352 (1989).
- Yamanaka, N. *et al.* An information linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. *DNA Res.* **8**, 61–72 (2001).
- Yamamoto, E., Karakaya, H. C. & Knap, H. T. Molecular characterization of two soybean homologues of *Arabidopsis* thaliana *CLAVATA1* from the wild type and fasciation mutant. *Biochem. Biophys. Acta.* **1491**, 333–340 (2000).
- Akao, S. & Kouchi, H. A supernodulating mutant isolated from soybean cultivar Enrei. *Soil. Sci. Plant Nutr.* **38**, 183–187 (1992).
- Matsubayashi, Y., Ogawa, M., Morita, A. & Sakagami, Y. An LRR receptor kinases involved in perception of a peptide plant hormone, phytosulfokine. *Science* **296**, 1470–1472 (2002).
- Scheer, J. M. & Ryan, C. A. The systemin receptor SR160 from *Lycopersicon peruvianum* is a member of the LRR receptor kinase family. *Proc. Natl Acad. Sci. USA* **99**, 9585–9590 (2002).
- Flemetakis, E. *et al.* *Lotus japonicus* contains two distinct *ENOD40* genes that are expressed in symbiotic, nonsymbiotic, and embryonic tissues. *Mol. Plant-Microbe Interact.* **13**, 987–994 (2000).

- Kouchi, H., Takane, K., So, R. B., Ladha, J. K. & Reddy, P. M. Rice ENOD40: isolation and expression analysis in rice and transgenic soybean root nodules. *Plant J.* **18**, 121–129 (1999).
- Kawaguchi, M. *et al.* Providing the basis for genomics in *Lotus japonicus*: the accessions Miyakojima and Gifu are appropriate crossing partners for genetic analyses. *Mol. Gen. Genomics* **266**, 157–166 (2001).
- Hayashi, M. *et al.* Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F2 population. *DNA Res.* **8**, 301–310 (2001).
- Kawasaki, S. & Murakami, Y. Genome analysis of *Lotus japonicus*. *J. Plant Res.* **113**, 497–506 (2000).
- Stiller, J. *et al.* High frequency transformation and regeneration of transgenic plants in model legume *Lotus japonicus*. *J. Exp. Bot.* **48**, 1357–1365 (1997).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank A. A. T. van den Brussel, J. Harper, P. M. Gresshoff and T. Ohyama for comments, and S. Sato, Y. Nakamura and S. Tabata for providing draft sequence data on TAC clones and for gene modelling. We also thank E. Kobayashi for horticultural services. This research was supported in part by Grants-in-Aid for Scientific Research on Priority Areas and Special Coordination Funds for Promoting Science and Technology to M.K. from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. This work was also partly supported by a grant from the programme for Promotion of Basic Research Activities for Innovative Biosciences to K.H.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K.
(e-mail: masayosi@env.sc.niigata-u.ac.jp). The sequences for BAC259.12D-1, BAC259.12D-2, *L. japonicus* *HAR1* gene, and soybean *NTS1* gene have been deposited in the DDBJ database under accession numbers AB092808, AB092809, AB092810 and AB092811, respectively.

Self-recognition promotes the foreign antigen sensitivity of naive T lymphocytes

Irena Štefanová^{*†}, Jeffrey R. Dorfman^{*†‡} & Ronald N. Germain^{*}

^{*} *Lymphocyte Biology Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA*

[†] *These authors contributed equally to this work*

Major histocompatibility complex (MHC) class I and II molecules are highly polymorphic proteins that bind and present foreign peptides to the clonally distributed $\alpha\beta$ receptors (TCR) of T lymphocytes. As a population, the immature T lymphocytes generated in the thymus express a very diverse set of TCR specificities. A process of positive selection filters this broad repertoire to optimize peripheral T cells for antigen recognition in the context of available MHC products. Only those precursor T cells whose TCRs generate an adequate but not excessive signalling response to self-peptides bound to the expressed MHC proteins undergo successful maturation¹. Here we show that post-thymic self-recognition facilitates the antigen reactivity of mature T cells. Both experimental and physiological interruption of T-cell contact with self-peptide MHC ligands leads to a rapid decline in signalling and response sensitivity to foreign stimuli. Because the adaptive immune system must be recruited early in an infectious process when antigen is limiting², these findings suggest that positive selection ensures predictable T-cell recognition of available self-ligands, which in turn promotes efficient responses to pathogens.

One of the consequences of positive selection is the maintenance of the TCRs of naive T cells in a state of partial phosphorylation,

[‡] Present address: Kenya Medical Research Institute, Wellcome Trust Research Laboratories, PO Box 230, Kilifi, Kenya.

owing to recognition of self-peptide MHC molecule ligands in peripheral lymphoid tissues^{3,4}. This continuous subthreshold signalling might enhance peripheral T-cell sensitivity to a low level of foreign antigen in a manner analogous to how weak signals facilitate response in other biological systems⁵. Alternatively, because a similar partial phosphorylation pattern is induced by inhibitory ligands of the TCR (antagonists), self-recognition by mature T cells could have an inhibitory influence that might control autoimmunity⁶. To distinguish between these divergent hypotheses, we examined the signalling and activation properties of naive T cells deprived of self-MHC contact. Removing TCR transgenic CD4⁺ T cells from secondary lymphoid tissues and placing them in dispersed culture at 37°C for as little as 15–30 min led to a rapid loss of pre-existing partial TCR ζ -chain phosphorylation (Fig. 1a) and substantial blunting of proliferation in response to foreign antigen compared with cells that were stimulated immediately after isolation (Fig. 1b). Control studies demonstrated that the depressive effects of pre-culture were not explained by a loss of viability among the T cells or by pre-sensitization to TCR-signalled death (Supplementary Fig. 1). This rapid loss of sensitivity in culture may explain the more extensive *in vivo* activation of p38 mitogen-activated protein kinase in TCR transgenic T cells by specific antigen, as compared with T cells stimulated *in vitro* after an isolation that included an exposure of cells to 37°C for 25 min⁷.

To examine the consequences of *in vivo* loss of self-recognition, TCR transgenic animals (H-2^b) were injected with monoclonal antibody to the MHC class II molecule involved in positive selection. Cells recovered 36 h after such an injection were stimulated with foreign antigen. As with the cells placed in dispersed culture at 37°C, the T cells deprived of self-recognition *in vivo* showed a substantial loss in TCR signalling sensitivity and functional antigen responsiveness (proliferation) in comparison with those recovered from mock-treated mice (Fig. 2a–d). Similar effects were seen as early as 8 h after antibody injection (data not shown). Signalling for an early activation response with a very low signalling threshold (CD69 upregulation⁸) was comparable for control cells and cells from animals treated with antibody (Fig. 2e). Furthermore, all self-ligand-deprived T cells underwent a cell size increase in response to agonist ligand that was comparable after 16 h to that of control cells (Supplementary Fig. 2b, c). Taken together, these results indicate that the MHC-deprived T cells retained some capacity for receptor signalling. The loss of antigen sensitivity induced by interference with self-recognition became detectable by cell size analysis after 31 h of incubation (Supplementary Fig. 2b, c), and was evident when the T cells were analysed for synthesis of the cytokine interleukin-2 (IL-2), a response that requires a higher intensity and more prolonged TCR signal⁹ (Fig. 2f). No evidence of enhanced spontaneous or TCR-induced death was seen among the cells from animals treated with antibody (Fig. 2g).

The preceding experiments did not address the role of MHC polymorphism in sustaining T-cell sensitivity. In addition, binding an antibody to MHC class II molecules on antigen-presenting cells (APCs) *in vivo* can potentially alter the physiology of the presenting cell itself or act by blocking signalling through the MHC class II-binding CD4 co-receptor¹⁰. Therefore, CD4⁺ TCR transgenic cells were transferred to recipients that either lacked detectable MHC class II molecule expression¹¹ or that expressed an MHC class II allele (H-2^q) distinct from that involved in the positive selection. The former tested the general role of MHC class II molecules in maintaining lymphocyte sensitivity, whereas the latter specifically examined the role of TCR self-recognition in distinction to the MHC class II–CD4 interaction. In both cases, the recovered T cells showed a similar loss of foreign antigen sensitivity relative to cells transferred into and recovered from an environment expressing the selecting MHC class II molecule (Fig. 2h). No signs of alloreactive responses of H-2^q CD8⁺ T cells were observed in those cultures lacking pigeon cytochrome *c* (PCC) peptide or in control co-

cultures of fresh AND T cells and H-2^q lymph node cells (Supplementary Fig. 3).

To determine whether these findings apply to a polyclonal population of T cells, β 2 microglobulin-deficient C57BL/6 mice with a normal population of CD4⁺ T cells but lacking most CD8⁺ T cells¹² were injected with blocking monoclonal antibody against MHC class II molecules. Recovered lymphocytes were then exposed to the bacterial superantigen staphylococcal enterotoxin A (SEA), which stimulates T cells expressing diverse TCRs¹³. The polyclonal CD4⁺ T cells from animals in which MHC class II recognition was blocked showed a marked loss of sensitivity to TCR stimulation as compared with cells from control mice (Fig. 2i, j). Similar results were obtained with staphylococcal enterotoxin B (SEB) (data not shown).

To place these observations in a physiological context, we examined an *in vivo* circumstance in which CD4⁺ T cells are temporarily without the opportunity for self-MHC molecule recognition; that is, during their transit between secondary lymphoid tissues by means of the blood¹⁴. Freshly isolated, blood-derived TCR transgenic T cells had nearly undetectable levels of partially phosphorylated TCR ζ and displayed lower sensitivity to foreign antigen compared with T cells from secondary lymphoid tissue (Fig. 3). This is the same biochemical and functional pattern seen with cells experimentally deprived of TCR interaction with self-peptide MHC ligands (Figs 1 and 2).

Several factors may contribute to the self-induced maintenance of antigen sensitivity. Subthreshold stimulation may promote cellular responses through increases in the local concentration of signalling molecules⁵, such as the modification/recruitment of various pro-activating components of the TCR signal transduction apparatus, including binding of the kinase ZAP-70 to partially phosphorylated TCR ζ chains. Furthermore, recognition of self-ligand may promote a particular spatial organization of TCRs that could facilitate signalling in response to stronger ligands, in analogy to the clustering of glutamate receptors induced by low-level spontaneous neurotransmitter release at neuromuscular junctions¹⁵. This clustering is essential for fast and reliable transmission of action potentials. We therefore examined the distribution of TCR complexes on T cells *in situ*. Immunofluorescent staining of lymph node sections revealed a crescent-shaped polarization of TCRs to the region of the lymphocyte membrane that was apposed to the MHC class II-rich membrane of interdigitating cells with a dendritic morphology (Fig. 4a). An asymmetric distribution of TCRs was also seen on most of the CD4⁺ T cells freshly isolated from lymph

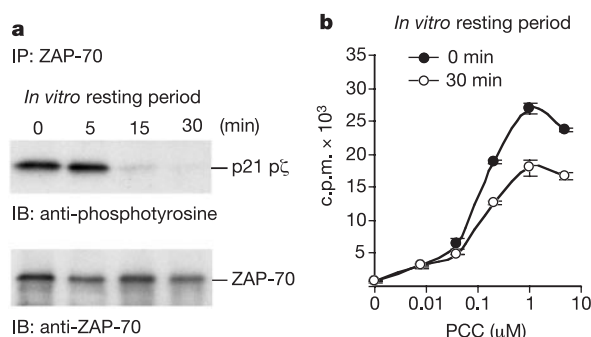


Figure 1 Acute loss of TCR ζ phosphorylation and antigen sensitivity by cultured naive T cells. **a**, **b**, Naive 5C.C7 CD4⁺ T cells were isolated from lymph nodes in ice-cold medium and either stimulated immediately or warmed to 37°C for varying times. The cells were lysed and the proteins subjected to immunoprecipitation (IP) and immunoblot (IB) analysis for tyrosine phosphorylation of the TCR ζ chain (**a**). Alternatively, freshly isolated or cultured cells were mixed with DCEK-presenting cells and varying concentrations of PCC peptide 88–104, then assayed at 60 h for proliferation using ³H-thymidine incorporation (**b**).

nodes (Fig. 4b–d). In agreement with our observations of self-ligand-sustained TCR ζ phosphorylation and signalling sensitivity, TCR polarization diminished within 30 min of *in vitro* culture at 37 °C (Fig. 4b–d), after injection of anti-MHC class II antibody *in vivo* (Fig. 4e, f), or following physiological loss of contact with MHC class II molecules by CD4⁺ T cells in the bloodstream (Fig. 4g).

These findings support the hypothesis that self-recognition maintains the sensitivity of T cells. They stand in contrast to predictions of the depression of T-cell responsiveness by self-ligand engagement that were based on an analogy to antagonist ligand function. To better understand why T cells behave differently in these two situations (self-recognition compared with antagonist exposure) despite similar patterns of TCR subunit phosphorylation,

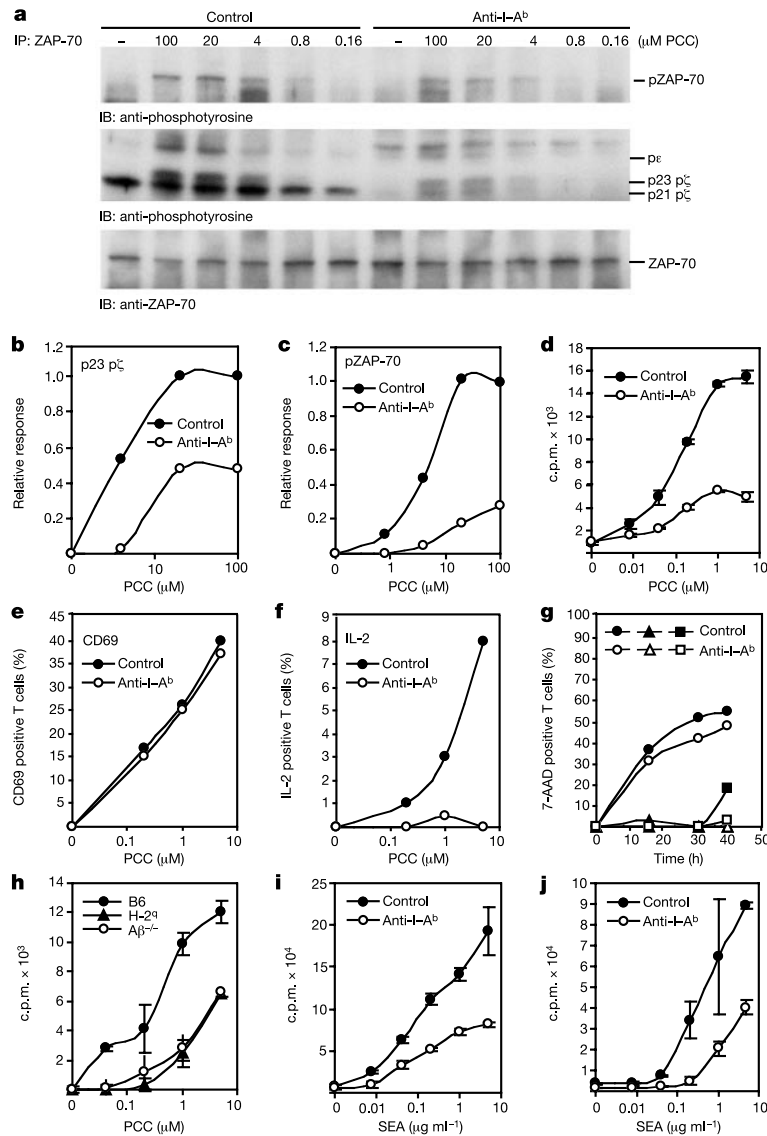


Figure 2 Marked decline in TCR signalling and sensitivity to foreign antigen among naive CD4⁺ T cells deprived of self-MHC class II recognition *in vivo*. **a–d**, AND H-2^b mice were either mock-injected or injected with monoclonal antibody Y3P specific for the peptide-binding domain of I-A^b—the only MHC class II molecule expressed in these animals and responsible for positive selection of the naive AND CD4⁺ T cells. **a**, Phosphotyrosine immunoblot analysis of the TCR proximal signalling response of freshly isolated cells from control and anti-I-A^b-treated animals across a range of offered PCC peptide using DCEK as presenting cells. These data are quantified in **b** (p23 phospho- ζ) and **c** (phospho-ZAP-70). Relative response was calculated as a ratio between the density for each experimental condition and the density of p23 phospho- ζ or phospho-ZAP-70 in control T cells exposed to APCs pulsed with 100 μM PCC. **d**, Proliferative responses of lymph node cells from control and anti-MHC class II antibody-treated animals stimulated with PCC peptide. **e–g**, Single cell analyses of T cells isolated from mock-injected and anti-I-A^b antibody (Y3P) injected AND H-2^b mice. T cells were placed in culture with DCEK APCs pre-pulsed with the indicated concentrations of PCC peptide, and analysed after 6 h

for cell surface expression of CD69 (**e**) and intracellular IL-2 (**f**). **g**, Cell viability was assayed over a 40-h period using 7-aminoactinomycin D. Circles, viability values of T cells incubated without added APCs; triangles, T cells cultured in the presence of APCs; squares, T cells stimulated in the presence of APCs pre-pulsed with 5 μM PCC. **h**, Naive CD4⁺ T cells from H-2^b AND donors were labelled with CFSE and transferred intravenously into wild-type H-2^b, Aβ^{-/-}, or wild-type H-2^q recipients. Thirty-six hours later, lymph nodes were recovered and equal numbers of unpurified transferred cells were placed in culture with I-E^k-expressing DCEK cells and various concentrations of PCC peptide. ³H-thymidine incorporation as an indicator of cell proliferation was measured 80 h later. **i**, **j**, β2m^{-/-} H-2^b mice were either mock-injected or injected with Y3P antibody to I-A^b to block MHC class II recognition. Thirty hours later, lymph node cells were placed in culture with either DCEK (**i**) or B10.A splenic presenting cells (**j**) and the indicated concentration of SEA. ³H-thymidine incorporation was measured 60 h later as an indication of the proliferative response.

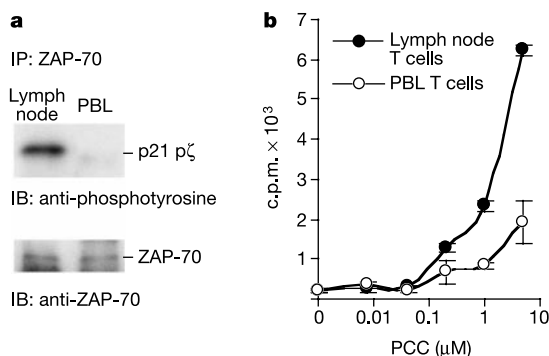


Figure 3 Peripheral blood CD4⁺ T cells lacking MHC class II contact have lower TCR ζ phosphorylation and reduced antigen sensitivity compared with lymph node cells from the same animals. **a**, **b**, Peripheral blood and lymph node CD4⁺ T cells from H-2^b AND mice were isolated and analysed by immunoprecipitation and phosphotyrosine immunoblotting for the state of the TCR ζ -chain phosphorylation (**a**), or placed in culture with B10.A splenic presenting cells and stimulated with PCC peptide for 60 h before analysis of proliferation using ³H-thymidine incorporation (**b**). Longer exposure of the blot in **a** revealed a low level of phosphorylated ζ -chains in anti-ZAP-70 immunoprecipitates from PBLs (data not shown).

we examined the distribution of the inhibitory tyrosine phosphatase SHP-1. This enzyme is recruited to TCRs engaging antagonist ligands¹⁶, where it interferes with proximal signalling. T cells exposed to antagonist-bearing APCs showed extensive membrane association of SHP-1 in the region of receptor polarization, whereas freshly isolated CD4⁺ T cells showed relatively little SHP-1 accumulation in the region of clustered TCRs (Fig. 4h, i). These results are concordant with the functional data reported above, as well as with evidence for differences between freshly isolated and antagonist-stimulated T cells in the specific sites of ζ -chain tyrosine phosphorylation¹⁷.

Positive selection of T cells in the thymus is a phenomenon recognized for nearly 25 years. Here we provide evidence for a previously unappreciated consequence of this process. Positive selection ensures that peripheral T cells express TCRs capable of functional binding to the pool of self-ligands available before introduction of a foreign ligand. In turn, this predictable TCR engagement actively sustains the naive T cells in an optimal state of sensitivity; that is, they are prepared to respond to a low density of foreign peptide MHC molecule ligand early in an infection². The rapid loss of this sensitive state on disengagement from self-MHC ligands suggests that it can be considered as a form of 'short-term memory'. Both biochemical modification of the TCR signalling

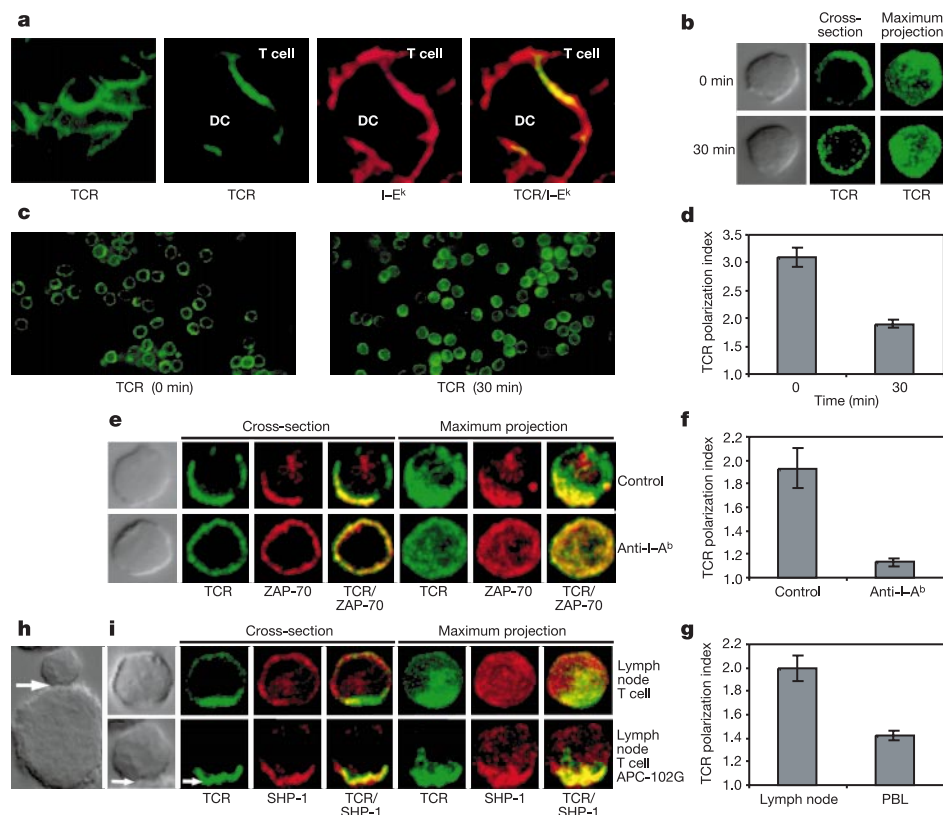


Figure 4 Naive CD4⁺ T cells show polarized TCR distributions that are maintained by contact with self-MHC class II molecules. **a**, Frozen sections of 5C.7 lymph nodes were stained for TCR β -chain (green) and I-E^k (red). The left panels show the TCR staining of a cluster of T cells. The right panels show the TCR and MHC class II staining of a single T cell-presenting cell pair. Co-localized TCR and MHC class II appear yellow in the superimposed image. DC, dendritic cells. **b–d**, 5C.7 CD4⁺ T cells were either fixed immediately after isolation or after incubation for 30 min at 37 °C and stained with anti-TCR β (**b**). Whole-field views of freshly isolated and *in vitro* rested T cells are shown in **c**. The frequency and degree of TCR polarization, expressed as the TCR polarization index (see Methods), is shown in **d**. **e**, **f**, Freshly isolated CD4⁺ T cells from control and anti-I-

A^b-treated AND mice were fixed, stained for either TCR β (green) or ZAP-70 (red) (**e**), and the TCR polarization index was calculated (**f**). **g**, CD4⁺ T cells from AND mice isolated from lymph nodes or blood were fixed, stained for TCR β , and the TCR polarization index was calculated. **h**, **i**, Naive CD4⁺ T cells from lymph nodes of 5C.7 transgenic mice were either fixed immediately or stimulated for 1 min with DCEK presenting cells pre-pulsed with 100 μ M antagonist peptide MCC (102G) and fixed. A full interference contrast view of the T cell-APC conjugate is shown in **h**. The cells were first stained for TCR (green), then permeabilized and stained for SHP-1 (red) (**i**). Arrows indicate contact between T cell and APC.

environment and the self-ligand-sustained polarization of the TCRs we report here may contribute to this effect. Concentration of TCRs increases the likelihood that contact with an APC bearing low densities of foreign ligand will result in a sufficient number of engagement events to trigger cell activation. Greater TCR density would also increase the effective on-rate of the binding reaction. Finally, receptor clustering can facilitate signal transduction, possibly by promoting biochemical crosstalk among nearby receptors^{18–20}. Induction of elevated levels of the inhibitory surface molecule CD5 (ref. 21) does not seem to override the positive effects of self-recognition. Peripheral blood CD4⁺ T cells show slightly lower CD5 levels than lymph node T cells from the same animals (data not shown), but are less sensitive despite this. In one report, CD4⁺ T cells deprived of self recognition *in vivo* did not show a difference in response compared to control cells³. However, the comparison involved cells subjected to a period of *in vitro* manipulation before stimulation, which may have led to a loss of sensitivity among those from the self-ligand-containing environment.

Evidence has emerged that MHC molecule binding can sustain memory CD4⁺ T cells in a state of heightened antigen reactivity²². Other studies have shown that co-presentation of 'null' ligands can augment T-cell responses at low densities of foreign ligand^{23,24}, raising the possibility that self-recognition may also contribute to T-cell activation concurrent with foreign antigen recognition. These data are in accord with the conclusion reached here, that self-recognition has a role in augmenting immunity rather than constraining it. □

Methods

Mice

All TCR transgenic mice were *Rag2*^{−/−} mice²⁵ that were homozygous for the TCR transgene. C57BL/6NCr Ly5.1 congenic mice (B6 H-2^b) were obtained from the National Cancer Institute. B6 *β2m*^{−/−} (ref. 12) and B10.D1(H-2^d) mice were purchased from Jackson Laboratories. B6 *Aβ*^{−/−} mice¹¹, H-2^b AND TCR transgenic²⁶ *Rag2*^{−/−} mice, and 5C.C7 TCR transgenic²⁷ *Rag2*^{−/−} mice were obtained from the National Institute of Allergy and Infectious Diseases contract colony maintained at Taconic Farms.

In vitro resting and antibody injections

For *in vitro* resting, cells were incubated at 37 °C for the indicated time in RPMI-1640 with 10% heat-inactivated fetal calf serum. For *in vivo* MHC class II blocking, mice were injected intraperitoneally with 1 to 2 mg of anti-I-A^b (clone Y3P²⁸) in two or three separate intraperitoneal injections over 30–48 h preceding recovery of lymph node cells. When TCR transgenic mice were used, splenocytes were recovered and combined with the lymph node cells. Use of anti-I-A^b in AND TCR transgenic mice prevents the antibody treatment from interfering with subsequent I-E^k-dependent functional assays.

Peripheral blood cell analysis

Peripheral blood lymphocytes (PBLs) were obtained from a retro-orbital bleed performed under anaesthesia in accordance with NIH veterinary standards. Blood was collected into heparinized Monoject tubes (Sherwood Medical). Erythrocytes were lysed using ACK lysing buffer (Quality Biological). MHC class II-expressing cells were removed from both PBLs and T cells collected from lymph nodes as described below to eliminate their potential contributions in functional experiments. To verify that the processing of PBL T cells does not influence their sensitivity status, T cells removed from lymph nodes were added into blood collected from T-cell-deficient donors (*CD3ε*^{−/−} B10.A), treated identically to PBL T cells, and their responsiveness compared to T cells freshly isolated from lymph nodes. Proliferative responses were comparable for both groups.

AND transgenic T-cell transfers

CD4⁺ T cells were enriched from a mixture of H-2^b AND TCR transgenic lymph node and spleen cells by negative selection at 4 °C using M450 sheep anti-rat immunoglobulin-γ (IgG) magnetic beads (DynaL Corp.) following incubation of the cells with a cocktail of antibodies that included anti-I-A^b (B21-2), anti-B220 (RA3-3A1), anti CD11c (N418), anti-CD11b (M1/70), anti-DEC-205 (NLDC-145) and anti-FcγRIII/III (2.4G2). Hybridoma cell culture supernatants were used except for purified M1/70, which was a gift from B. Kelsall and A. Iwasaki. Enriched CD4⁺ T cells were dye-labelled for 5 min with 0.3 μM 5-(and 6-) carboxyfluorescein diacetate succinimidyl ester (CFDA-SE or CFSE; Molecular Probes), washed, and injected intravenously into the indicated recipients. Recipients were pretreated with 300–500 μg purified anti-NK1.1 (PK136). Twenty-four to thirty-six hours after transfer, lymph node cells were recovered from the recipients. Recovered lymph node cells containing comparable numbers of transferred, labelled AND T cells per assay well were stimulated *in vitro* as described below without further purification.

Antigen/superantigen stimulation and proliferation

Stimulator cells were I-E^k-expressing, B7.1^{negative-low}, ICAM-1^{negative} fibroblasts (DCEK²⁹) or I-E^k-expressing B10.A splenocytes, exposed to 3,000 rad γ-irradiation before use. Comparable results were obtained using either type of presenting cell. The stimulator cells were pre-incubated for 4 h with the indicated concentration of peptide (PCC 88–104) or staphylococcal enterotoxin A (SEA, Toxin Technologies). For B6 *β2m*^{−/−} responders, lymph node cells were used unpurified as responders. T cells and washed stimulator cells were incubated together for 48–72 h (TCR transgenic T cells, 5 × 10⁴ per well except for cell transfers) or 72–120 h (polyclonal T cells, 1 × 10⁵ per well and cell transfer experiments, 0.5–1 × 10⁴ AND T cells per well) in 96-well plates. A total of 1 or 2 μCi ³H-thymidine (New England Nuclear) was added to each well, and cultures were incubated for a further 8–15 h before incorporation was measured.

Single cell response analysis

The T cells were stimulated with DCEK cells pre-pulsed with PCC peptide and analysed after 6 h for CD69 and IL-2 expression³⁰, and at the indicated time points for viability using 1-aminoactinomycin D staining (Sigma).

Immunoprecipitation and immunoblot analysis

For signalling analysis, DCEK stimulator cells were used and cells were stimulated for 1 min before pelleting and lysis. Cells were lysed in buffer containing 1% NP-40 (Pierce) with protease and phosphatase inhibitors. ZAP-70 precipitation and protein or phosphotyrosine immunoblotting were performed as described¹⁶.

Immunofluorescence microscopy

For immunohistochemistry, 17-μm frozen sections of lymph node were used. For immunofluorescence staining, freshly isolated T cells were first fixed in 4% paraformaldehyde in PBS for 5 min at 37 °C. Antibodies, fluorochromes, staining procedures, and methods for confocal analysis and quantification are provided in the Supplementary Information.

Received 2 April; accepted 4 September 2002; doi:10.1038/nature01146.

1. Robey, E. & Fowlkes, B. J. Selective events in T cell development. *Annu. Rev. Immunol.* **12**, 675–705 (1994).
2. Cohn, M. & Langman, R. E. The protecton: the unit of humoral immunity selected by evolution. *Immunol. Rev.* **115**, 11–147 (1990).
3. Witherden, D. *et al.* Tetracycline-controllable selection of CD4(+) T cells: half-life and survival signals in the absence of major histocompatibility complex class II molecules. *J. Exp. Med.* **191**, 355–364 (2000).
4. Dorfman, J. R., Štefanová, I., Yasutomo, K. & Germain, R. N. CD4 + T cell survival is not directly linked to self-MHC-induced TCR signaling. *Nature Immunol.* **1**, 329–335 (2000).
5. Taniguchi, T. & Takaoka, A. A weak signal for strong responses: interferon-α/β revisited. *Nature Rev. Mol. Cell Biol.* **2**, 378–386 (2001).
6. Kersh, B. E., Kersh, G. J. & Allen, P. M. Partially phosphorylated T cell receptor ζ molecules can inhibit T cell activation. *J. Exp. Med.* **190**, 1627–1636 (1999).
7. Zell, T. *et al.* Single-cell analysis of signal transduction in CD4 T cells stimulated by antigen *in vivo*. *Proc. Natl Acad. Sci. USA* **98**, 10805–10810 (2001).
8. Hemmer, B., Štefanová, I., Vergelli, M., Germain, R. N. & Martin, R. Relationships among TCR ligand potency, thresholds for effector function elicitation and the quality of early signaling events in human T cells. *J. Immunol.* **160**, 5807–5814 (1998).
9. Iezzi, G., Karjalainen, K. & Lanzavecchia, A. The duration of antigenic stimulation determines the fate of naive and effector T cells. *Immunity* **8**, 89–95 (1998).
10. Doyle, C. & Strominger, J. L. Interaction between CD4 and class II MHC molecules mediates cell adhesion. *Nature* **330**, 256–259 (1987).
11. Crump, A. L., Grusby, M. J., Glimcher, L. H. & Cantor, H. Thymocyte development in major histocompatibility complex-deficient mice: evidence for stochastic commitment to the CD4 and CD8 lineages. *Proc. Natl Acad. Sci. USA* **90**, 10739–10743 (1993).
12. Koller, B. H., Marrack, P., Kappler, J. W. & Smithies, O. Normal development of mice deficient in β2M, MHC class I proteins, and CD8 + T cells. *Science* **248**, 1227–1230 (1990).
13. Gascoigne, N. R. Interaction of the T cell receptor with bacterial superantigens. *Semin. Immunol.* **5**, 13–21 (1993).
14. Butcher, E. C., Williams, M., Youngman, K., Rott, L. & Briskin, M. Lymphocyte trafficking and regional immunity. *Adv. Immunol.* **72**, 209–253 (1999).
15. Saitoe, M., Schwarz, T. L., Umbach, J. A., Gundersen, C. B. & Kidokoro, Y. Absence of junctional glutamate receptor clusters in *Drosophila* mutants lacking spontaneous transmitter release. *Science* **293**, 514–517 (2001).
16. Dittel, B. N., Štefanová, I., Germain, R. N. & Janeway, C. A. Jr Cross-antagonism of a T cell clone expressing two distinct T cell receptors. *Immunity* **11**, 289–298 (1999).
17. Kersh, E. N., Shaw, A. S. & Allen, P. M. Fidelity of T cell activation through multistep T cell receptor ζ phosphorylation. *Science* **281**, 572–575 (1998).
18. Bray, D., Levin, M. D. & Morton-Firth, C. J. Receptor clustering as a cellular mechanism to control sensitivity. *Nature* **393**, 85–88 (1998).
19. Germain, R. N. & Štefanová, I. The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation. *Annu. Rev. Immunol.* **17**, 467–522 (1999).
20. Gestwicki, J. E. & Kiessling, L. L. Inter-receptor communication through arrays of bacterial chemoreceptors. *Nature* **415**, 81–84 (2002).
21. Smith, K. *et al.* Sensory adaptation in naive peripheral CD4 T cells. *J. Exp. Med.* **194**, 1253–1261 (2001).
22. Kassiotis, G., Garcia, S., Simpson, E. & Stockinger, B. Impairment of immunological memory in the absence of MHC despite survival of memory T cells. *Nature Immunol.* **3**, 244–250 (2002).
23. Micheletti, F. *et al.* Supra-agonist peptides enhance the reactivation of memory CTL responses. *J. Immunol.* **165**, 4264–4271 (2000).
24. Wülfing, C. *et al.* Costimulation and endogenous MHC ligands contribute to T cell recognition. *Nature Immunol.* **3**, 42–47 (2002).

25. Shinkai, Y. *et al.* Restoration of T cell development in RAG-2-deficient mice by functional TCR transgenes. *Science* **259**, 822–825 (1993).
26. Kaye, J. *et al.* Selective development of CD4 + T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341**, 746–749 (1989).
27. Seder, R. A., Paul, W. E., Davis, M. M. & Fazekas de St Groth, B. The presence of interleukin 4 during *in vitro* priming determines the lymphokine-producing potential of CD4 + T cells from T cell receptor transgenic mice. *J. Exp. Med.* **176**, 1091–1098 (1992).
28. Janeway, C. A. Jr *et al.* Monoclonal antibodies specific for Ia glycoproteins raised by immunization with activated T cells: possible role of T cell bound Ia antigens as targets of immunoregulatory T cells. *J. Immunol.* **132**, 662–667 (1984).
29. Ding, L., Linsley, P. S., Huang, L.-Y., Germain, R. N. & Shevach, E. M. IL-10 inhibits macrophage costimulatory activity by selectively inhibiting the up-regulation of B7 expression. *J. Immunol.* **151**, 1224–1234 (1993).
30. Itoh, Y. & Germain, R. N. Single cell analysis reveals regulated hierarchical T cell antigen receptor signaling thresholds and intracellular heterogeneity for individual cytokine responses of CD4 + T cells. *J. Exp. Med.* **186**, 757–766 (1997).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank A. Fox for technical support, O. Schwartz for assistance with microscopy and image preparation, S. Stoll for assistance with mice, A. Iwasaki and B. Kelsall for reagents, and the members of the Germain laboratory for discussion.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.N.G.
(e-mail: rgermain@nih.gov).

HIV-1 superinfection despite broad CD8⁺ T-cell responses containing replication of the primary virus

Marcus Altfeld^{*,†}, Todd M. Allen^{*,†}, Xu G. Yu^{*}, Mary N. Johnston^{*}, Deepak Agrawal^{*}, Bette T. Korber[‡], David C. Montefiori[§], David H. O'Connor^{||}, Ben T. Davis^{*}, Paul K. Lee^{*}, Erica L. Maier^{*}, Jason Harlow^{*}, Philip J. R. Goulder[¶], Christian Brander^{*}, Eric S. Rosenberg^{*} & Bruce D. Walker^{*}

^{*} Partners AIDS Research Center and Infectious Disease Division, Massachusetts General Hospital and Division of AIDS, Harvard Medical School, Boston, Massachusetts 02129, USA

[‡] Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

[§] Department of Surgery, Duke University Medical Center, Durham, North Carolina 27710, USA

^{||} Wisconsin Regional Primate Research Center, University of Wisconsin, Madison, Wisconsin 53715, USA

[¶] Department of Pediatrics, Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, Oxford OX3 7LJ, UK

[†] These authors contributed equally to this work

Early treatment of acute HIV-1 infection followed by treatment interruptions has shown promise for enhancing immune control of infection^{1–3}. A subsequent loss of control, however, allows the correlates of protective immunity to be assessed. Here we show that sudden breakthrough of plasma viraemia occurred after prolonged immune containment in an individual infected with HIV-1 at a time when 25 distinct CD8⁺ T-cell epitopes in the viral proteins Gag, RT, Integrase, Env, Nef, Vpr, Vif and Rev were being targeted. Sequencing of the virus in plasma and cells showed that superinfection with a second clade-B virus was coincident with the loss of immune control. This sudden increase in viraemia was associated with a decline in half of the CD8⁺ T-cell responses. The declining CD8⁺ T-cell responses were coupled with sequence changes relative to the initial virus that resulted in impaired recognition. Our data show that HIV-1 superinfection can occur

in the setting of a strong and broadly directed virus-specific CD8⁺ T-cell response. The lack of cross-protective immunity for closely related HIV-1 strains, despite persistent recognition of multiple CD8 epitopes, has important implications for public health and vaccine development.

Increasing evidence suggests that virus-specific CD8⁺ T-cell responses are important in the control of HIV-1 or SIV replication^{4–8}, and inducing and maintaining these responses are considered to be central for the development of effective HIV-1 vaccines. In the simian-HIV (SHIV) model, induction of strong SHIV-specific CD8⁺ T-cell responses does not prevent infection after challenge but is associated with persistent control of low levels of viral replication and attenuated disease^{9–11}. Although similar vaccine regimens have failed to control more pathogenic and less easily neutralizable SIV strains^{12–14}, these studies nevertheless provide optimism that an effective vaccine to prevent HIV-1 disease progression in humans could be developed. Indeed, the augmentation of HIV-1-specific immunity by the early treatment of persons with acute HIV-1 infection, followed by supervised treatment interruptions (STI), has been shown^{1–3}. Most individuals treated by this method have been able to control, at least transiently, viraemia after the cessation of treatment at levels that did not require the re-initiation of treatment³. One such individual effectively controlled infection for over 290 d after STI but then experienced a marked increase in plasma viral load (Fig. 1a), offering us the opportunity to investigate the immunological and virological

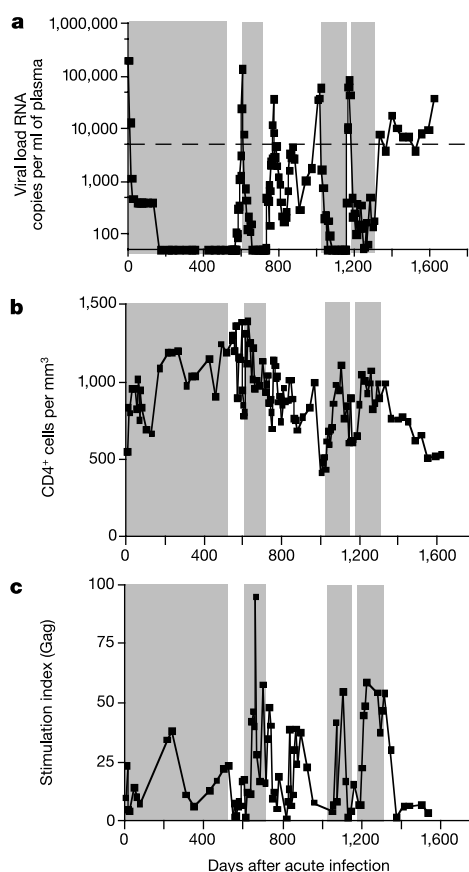


Figure 1 HIV-1 viral loads, CD4⁺ T-cell counts and Gag-specific lymphoproliferative responses in study subject AC-06. **a**, HIV-1 plasma viral loads; **b**, CD4⁺ T-cell counts; **c**, Gag-specific lymphoproliferative responses. Time zero represents the first presentation of AC-06 with symptomatic acute HIV-1 infection. At this time, HIV-1 viral loads were 8.8×10^6 RNA copies per ml of plasma (not shown). Shaded areas represent treatment with highly active antiretroviral therapy (HAART). Dotted line indicates 5,000 copies of HIV-1 RNA per ml of plasma.

correlates of loss of immune control.

Subject AC-06 underwent STI 546 d after early treatment of acute HIV-1 infection by a previously reported protocol³. The virus rebounded, and within 60 d (day 606 of infection) exceeded 50,000 HIV-1 RNA copies per ml of plasma, necessitating the re-initiation of therapy (Fig. 1a). Therapy was again interrupted at day 715. Although there was a transient peak in viraemia at day 773, viraemia declined spontaneously and his CD4⁺ T-cell counts remained stable. He then did not meet the criteria for re-initiating therapy until day 1,020 (305 d after stopping therapy), showing prolonged control of viraemia while off therapy.

At this stage, however, he experienced a sudden rapid increase in viraemia along with a marked drop in CD4⁺ T-cell counts from 998 cells per μ l (day 972) to 515 cells per μ l before HAART was reinitiated (Fig. 1b). Subsequent STI 4 months later (day 1,149) was associated with an even more rapid loss of control, in which viraemia exceeded 50,000 RNA copies per ml of plasma within 20 d. After another 4 months of highly active antiretroviral therapy (HAART), treatment was again stopped (day 1,312) and HIV-1-viraemia once more rebounded but to lower peak concentrations. Although viraemia never exceeded 50,000 RNA copies per ml of

plasma, secondary criteria to restart therapy were met at day 86 owing to viral loads above 5,000 RNA copies per ml for three consecutive weeks. At this stage, subject AC-06 chose to decline further drug therapy, with viral loads ranging from 3,700 to 18,200 RNA copies per ml for 280 d, before increasing to 38,000 RNA copies per ml on day 1,622. During this time, his CD4⁺ T-cell counts declined progressively.

We carried out comprehensive evaluation of the virus-specific immune responses to define the correlates of this loss of immune control. Gag-specific CD4⁺ T-cell proliferative responses were detected after early treatment^{3,15,16} and were present during the period of control during the second STI, but they declined with the sudden increase in viraemia (Fig. 1c). Although impairment of HIV-1-specific T-helper cell (T_H) responses by continuing low-level viral replication¹⁷ before the rebound of viraemia cannot be excluded by these data, they suggest that a sudden loss of Gag-specific T_H responses did not precede the observed loss of viral control in this individual. Similarly, only weak neutralizing antibody responses (titres ranging from <4 to 8 reciprocal plasma dilutions) against autologous virus were present during the second STI at the time of viral control.

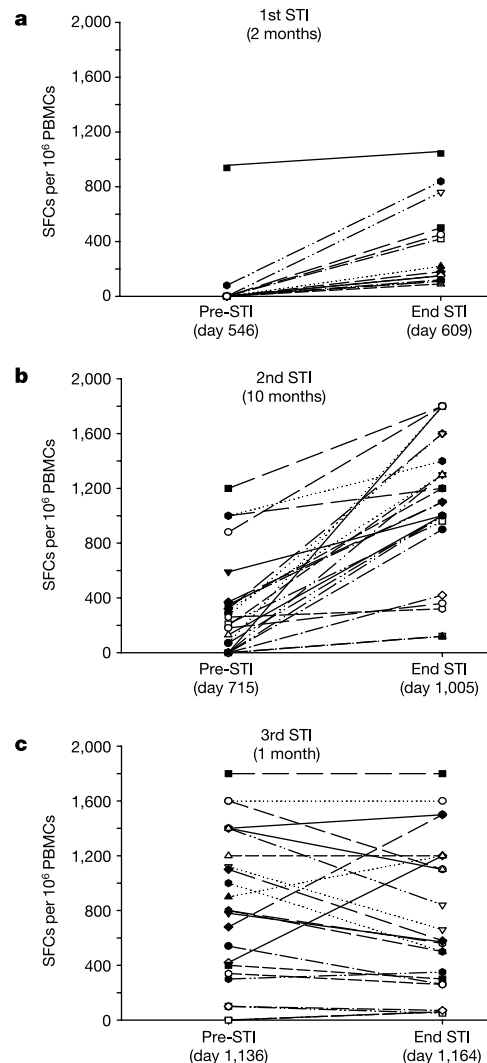


Figure 2 CD8⁺ T-cell responses to described optimal clade-B sequence cytotoxic T lymphocyte (CTL) epitopes during successive supervised treatment interruptions (STIs). Responses to individual epitopes are shown before and at the end of each STI (**a**, 1st STI; **b**, 2nd STI; **c**, 3rd STI) and are presented as spot-forming cells (SFCs) per million

peripheral blood mononuclear cells (PBMCs). The duration of STI is indicated in months. All epitope-specific responses increased until the third STI, during which more than half of the responses declined.

Because animal model data suggest that CD8⁺ T-cell responses are essential for immune control in AIDS virus infections^{5,7,18}, we measured longitudinally the maintenance of virus-specific CD8⁺ T-cell responses using an interferon- γ (IFN- γ) Elispot assay and 504 clade-B overlapping peptides spanning all expressed HIV-1 proteins. HIV-1-specific CD8⁺ T-cell responses were enhanced and broadened with re-exposure to antigen during the first and second STIs (see also ref. 19), ultimately targeting 25 distinct cytotoxic T lymphocyte (CTL) epitopes with a total of 27,470 spot-forming cells (SFCs) per 10⁶ peripheral blood mononuclear cells (PBMCs) at the end of the second STI (Fig. 2a, b). In addition to IFN- γ secretion after antigenic stimulation, these epitope-specific CD8⁺ T cells also demonstrated epitope-specific cytotoxic activity (data not shown). But this pattern of enhancement of epitope-specific responses present during the first two STIs changed markedly during the third STI. Although a subset of epitope-specific CD8⁺ T-cell

responses was maintained or expanded further, for the first time some established responses declined (Fig. 2c), leaving a reduced magnitude of 18,250 SFCs per 10⁶ PBMCs ($P = 0.01$, versus the second STI).

To investigate the relationship between sequence variation within targeted epitopes and loss of immune control¹⁸, we sequenced regions in HIV-1 *gag* using PBMC samples derived from the time of acute infection (day 18) and the third STI (day 1,170). Although sequence changes were identified in three of the seven epitopes located within HIV-1 *gag*, there was an even greater number of amino acid changes in the flanking regions that did not contain targeted CD8⁺ T-cell epitopes (Fig. 3a). Phylogenetic analysis of autologous *gag* sequences from these time points (day 18 acute infection, and day 1,170 3rd STI) indicated that a distinct and unrelated second strain of HIV-1 had emerged (Fig. 3b). Additional sequencing of virus derived from intermediate time points during

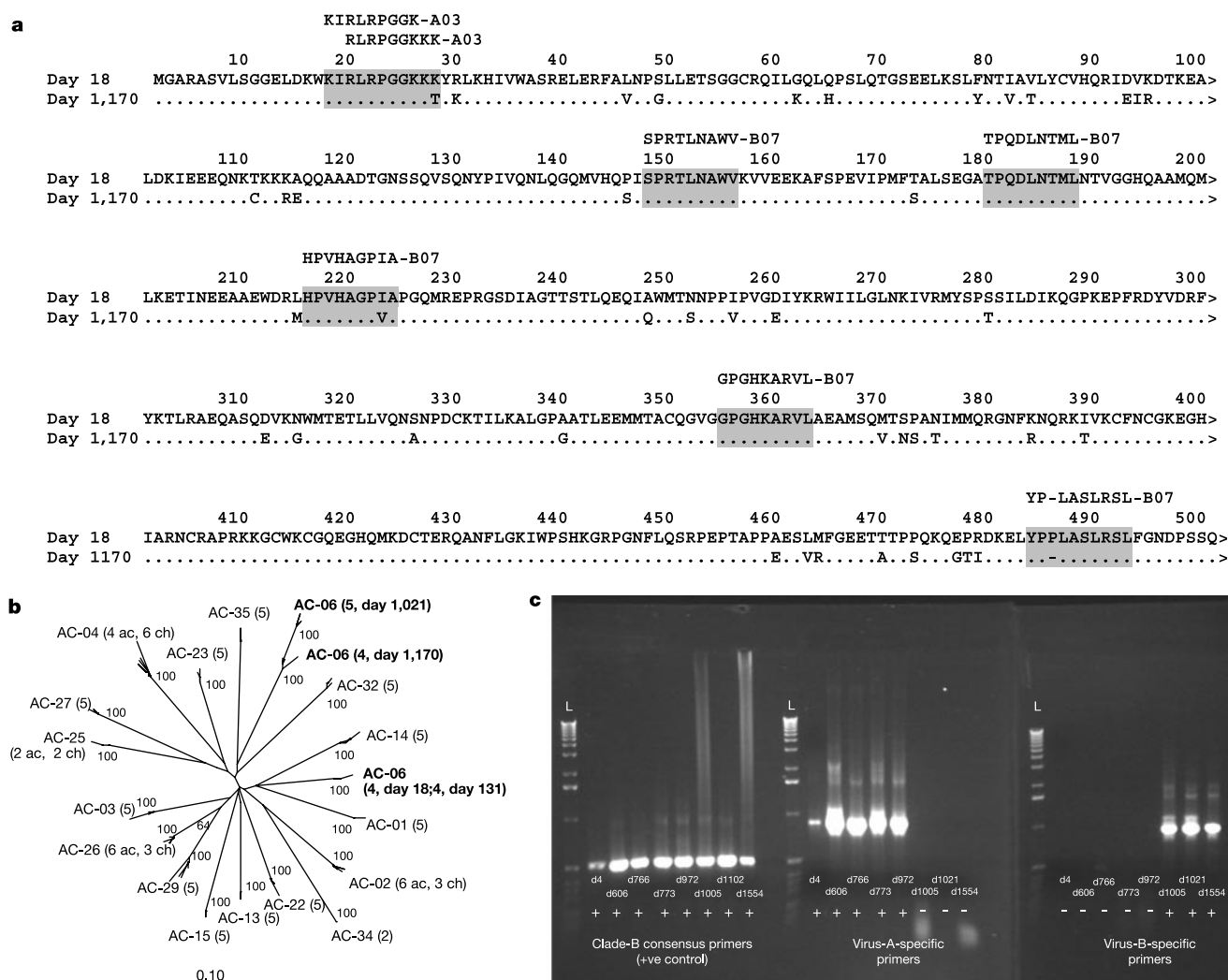


Figure 3 Sequence and phylogenetic analysis of viruses from subject AC-06. **a**, Gag sequences from subject AC-06 show high amino acid diversity. Consensus sequences were derived from several clones (minimum 4, mean 8) spanning each of the individual PCR fragments of the HIV-1 *gag* open reading frame. The sequences shown reflect the predominant viral populations present at day 18 and day 1,170. Boxes indicate the seven HLA-A03- or B07-restricted Gag epitopes targeted by this subject. **b**, HIV *gag* sequences from subject AC-06 during acute (day 18) and chronic (day 1,170) infection cluster independently. Sequences from 17 HIV-1 infected persons from the Boston area were aligned²⁷. The neighbour-joining phylogenetic tree indicates that two distinct viruses are present in subject AC-06 at these different times after infection (bootstrap values >60 are

shown). Numbers in parentheses indicate the number of sequences included for each subject and where available the number of acute (ac) or chronic (ch) stage sequences included. Scale bar indicates the genetic distance along the branches. **c**, Virus B predominates in plasma of subject AC-06 on day 1,005. Specific PCR primers for virus A and virus B were generated based on unique nucleotide substitutions present in the two viruses in a region of *gag*. RT-PCR was carried out on plasma viral RNA isolated throughout infection. Consensus clade-B-specific primers successfully amplify products at all time points, verifying sample integrity. Only virus-A-specific primers amplify product from plasma samples taken during the first 972 d of infection; by contrast, only virus-B-specific primers amplify product from samples taken from day 1,005 onwards.

the first STI (day 606) and the second STI (day 1,021) confirmed these results, which were verified by an independent investigator (data not shown). These data showed that the early samples (day 18 and day 606) represented one strain of HIV-1 (referred to as virus A), whereas later samples (day 1,021 and day 1,170) represented a distinct second strain of HIV-1 (referred to as virus B) that became detectable during the end of the second STI and persisted during the third and fourth STI.

Overall, the sequences of virus A and virus B were significantly different, showing 12% heterogeneity in amino acid sequence (on the basis of >80% sequence coverage of whole HIV genomes). BLAST searches of the HIV database also confirmed the uniqueness of both viruses. By contrast, viruses A and B themselves were found to vary only minimally (<0.3% amino acid heterogeneity) over time between day 18 and day 606 (virus A) and between day 1,021 and day 1,170 (virus B). Similarly, the viral diversity observed to develop in other subjects undergoing STIs over a similar time frame was also significantly more limited (data not shown). Thus, the sequence variations observed between virus A and virus B were not the result of viral evolution in several targeted CD8⁺ T-cell epitopes, but rather the emergence of a new HIV-1 clade-B strain in the study subject at a time when the initial virus was well controlled. An additional maximum likelihood phylogenetic analysis showed that it is highly unlikely that virus A and virus B arose from the same lineage within the context of the Boston cohort (see Supplementary Fig. 1). In addition, there was no evidence for recombination

between virus A and virus B in subject AC-06 (data not shown). Both strains were most closely related to HIV-1 clade-B sequences and possessed an R5 phenotype as determined by their differential infection in U87-CD4-CCR5, U87-CD4-CXCR4 and MT2 cells (ref. 20 and data not shown).

The above data suggest that either dual infection with viruses A and B occurred during the primary HIV-1 infection or there was subsequent superinfection with virus B. Review of medical records indicated that subject AC-06 experienced symptoms including fever, lymphadenopathy and night sweats in the month preceding the first detection of virus B. In addition, this subject reported an unprotected sexual risk exposure with an anonymous partner 4–8 weeks before the beginning of symptoms. The clinical data were thus consistent with exposure to another virus. To define further the time of infection with virus B, virus-specific nested polymerase chain reaction (PCR) primer sets corresponding to unique *gag* nucleotide sequences were used to track both viruses throughout infection (see Supplementary Table 1). Nested PCR with reverse transcription (RT-PCR) of plasma virus showed that virus A was the only detectable virus at the time of acute infection (day 4), during the first STI (day 606), and throughout most of the second STI (until day 972). In marked contrast, after the sudden increase in viraemia (day 1,005, Fig. 1a), the only virus detected in plasma was virus B (Fig. 3c). We carried out a similar analysis on viral DNA derived from PBMCs. Primers specific for virus A successfully amplified products from PBMC samples taken throughout infec-

Table 1 Viral sequence analysis and autologous epitope recognition by subject AC-06

Time Circulating virus				1st STI		2nd STI		3rd STI		4th STI	
				Beginning	End	Beginning	End	Beginning	End	Beginning	End
				Virus A	Virus A	Virus A	Virus B	Virus B	Virus B	Virus B	Virus B
Conserved and/or persistently recognized epitopes*											
1	A3-GK9 Pol	Virus A	GIPHPAGLK	0	0	0	1,500	790	1,500	670	1,000
		Virus B	-----								
2	A3-KK10 Pol	Virus A	KLVDFFRELNK	0	0	70	1,500	860	1,500	1,000	1,500
		B	-----								
3	B7-SV9 Gag	Virus A	SPRTLNAWV	0	0	0	120	30	120	30	280
		B	-----								
4	B7-TL9 Gag	Virus A	TPQDLNMTL	0	0	0	1,300	610	840	720	1,000
		B	-----								
5	B7-GL9 Gag	Virus A	GPGHKARVL †	480	1,200	1,000	1,500	1,500	1,500	1,000	1,500
		B	-----								
6	B7-SM9 Pol	Virus A	SPAIFQSSM	0	90	130	1000	750	690	590	500
		B	-----								
7	A3-KK9 Gag	Virus A	KIRLRPGGK	0	220	370	1100	480	390	270	150
		B	-----								
8	A3-RK10 Vlf	Virus A	RIRTWKSLVK	0	180	270	1500	350	580	330	900
		B	--S-----	0	100	250	300	200	650	360	820
9	A3-RK11 Vlf	Virus A	RRKPPPLPSIAK	0	150	180	360	150	650	340	520
		B	-T-----VT-	0	150	190	380	170	380	380	570
Variable epitopes and declining responses											
1	A3-QR9 Pol	Virus A	QIYAGIKVK	50	1,000	1,500	1,500	440	800	420	320
		B	-----R	30	760	1,000	550	130	130	140	40
2	A3-RK11 Pol	Virus A	RTRGAHTNDVK	0	0	0	860	850	260	130	120
		B	-----R	0	0	0	270	150	120	70	60
3	B7-HM0 Vlf	Virus A	HPRISSEVHI	0	270	380	1,200	1,000	800	630	370
		B	--K-----	0	270	320	1,200	1,000	800	800	360
4	A3-RK9 Ga9	Virus A	RLRPGGKKK	0	420	590	1,000	300	440	280	120
		B	-----T	0	0	0	380	80	170	60	50
5	B7-FL9 Vpr	Virus A	FPRTWLHGL	70	180	270	1000	520	430	280	300
		B	--W-----	0	0	80	220	70	210	160	140
6	A3-AK9 Pol	Virus A	AIFQSSMTK	0	0	0	420	100	70	0	40
		B	-----I-	0	0	0	40	0	0	0	0
7	A3-AK10 Pol	Virus A	AVFIHNFKRK	0	0	0	120	0	50	0	0
		B	---V-----	0	0	0	0	0	0	0	0
Responses specific to virus B											
1	B7-HA9 Gag	Virus A	HPVHAGPIA	0	0	0	120	110	140	130	640
		B	-----V-	0	0	0	130	140	190	160	900
2	B7-R10 Env	Virus A	RPSNNTRKSI	0	0	0	240	260	270	220	420
		B	--N-----RG-	0	0	0	360	230	1,000	650	1,200
3	B7-YL9 Gag	Virus A	YPP ^L ASLRSL	0	0	0	0	0	0	0	0
		B	-----	0	0	0	0	0	660	520	760

*Sequences were generated from cell-associated DNA isolated on day 18 (Virus A) and day 1,170 (Virus B).

†In the earliest sample sequenced (day 18), the autologous sequence for this epitope was GPGHKARVL, which was equally well recognized, whereas the GPGHKARVL sequence dominated by day 606.

Dot indicates position of amino acid infection in virus A.

tion, albeit less consistently from later time points when virus B predominated in plasma (data not shown). By contrast, primers specific for virus B successfully amplified products only from PBMC samples taken after day 1,021 (data not shown). Thus, the clinical and PCR data are most consistent with superinfection rather than with dual infection at the time of primary infection.

We next compared the sequence integrity of the CD8⁺ T-cell epitopes that were targeted before emergence of virus B to assess the impact of viral variability on immune recognition. Autologous sequences of virus A and virus B were amplified successfully for all regions except portions of *env* and *nef*, for which high sequence diversity limited amplification with the available primers (Supplementary Fig. 2 and Table II). Sixteen of the twenty-five epitopes were sequenced, as well as three new epitopes targeted for the first time after emergence of virus B. Recognition of autologous epitopes was assessed using PBMCs from several time points before and after emergence of virus B (Table 1) using an IFN- γ Elispot assay.

For 9 of the 16 sequenced epitopes targeted at the time of loss of immune control, either the epitope sequence was conserved completely (7 epitopes) or the virus B variant differed but was equally well recognized by CD8⁺ T cells derived from early and late time points (two epitopes), as assessed through peptide titration Elispot assays (ref. 21 and Supplementary Fig. 3). CD8⁺ T-cell responses to eight of these nine epitopes persisted after emergence of virus B (Table 1). The exception was a declining HLA-A3-restricted response to p17 Gag (KIRLRPGGK); this sequence was identical in both viruses, but virus B had an arginine to threonine mutation at position C + 2 that may have impaired antigen processing²².

For 7 of the 16 epitopes evaluated by sequence analysis, there were sequence changes in the CD8⁺ T-cell epitopes that were associated with declining CD8⁺ T-cell responses (Table 1) and decreased recognition by the established immune response (Supplementary Fig. 3). Indeed, comparison of the variability observed within each of these 16 sequenced epitopes to clade-B sequences present in the Los Alamos database showed a similar trend, with those epitopes conserved between virus A and virus B showing a greater degree of conservation than those that varied. Sequence data were not obtained for nine epitopes, for which five responses declined and four persisted (data not shown). Notably, the immunodominant CD8⁺ T-cell epitope GPGHKARVL in p24 Gag was present in both viruses, indicating that maintained responses against this epitope were not sufficient to prevent emergence of virus B^{9,18}.

Optimal CD4⁺ T-cell responses in this individual have not been defined and thus the impact that this viral diversity has on maintenance of HIV-specific CD4⁺ T-cell responses is unknown. After emergence of virus B, however, Gag-specific CD4⁺ T-cell proliferative responses did rebound to significant levels, suggesting persistent cross-recognition. Notably, a marked decline in Gag-specific CD4 proliferative responses and CD4⁺ T-cell counts was observed during the fourth STI in the setting of increased viral replication (Fig. 1). With respect to humoral responses, as stated previously, only weak neutralizing antibody responses specific to virus A were detectable before superinfection. But these virus-A-specific responses were lost after superinfection with virus B (titre < 4) and were not crossreactive against virus B (Supplementary Table III). After emergence of virus B, virus-B-specific neutralizing antibody responses developed for the first time, with titres reaching 40. But virus B subsequently escaped from these virus-B-specific neutralizing antibody responses (titre < 4). These data show that neutralizing antibodies to virus B were lacking at the time of superinfection but were produced later, potentially contributing to viral evolution and neutralization escape.

In addition to a loss of immune responses at the time of emergence of virus B, at least three new CD8⁺ T-cell responses within HIV-1 Gag and Env were detectable for the first time after superinfection, including two previously described CD8 epitopes and one newly described epitope in HIV-1 Gag. Fine mapping of the

new response showed that it was directed against an HLA-B7-restricted epitope (YPLASLRSL). Comparison of virus A and B sequences in this region indicated that the insertion of a proline at position 3 of this epitope (YPPLASLRSL) in virus A resulted in a variant that apparently had prevented the initial presentation of this region to the immune system (Fig. 3a and Supplementary Fig. 3). A response to this region was detectable for the first time when virus B predominated (Table 1). Similarly, for two additional epitopes that were recognized for the first time with emergence of virus B, the autologous variants of virus B were recognized at lower peptide concentrations than were the virus A variants (Supplementary Fig. 3). Thus, new CD8⁺ T-cell responses were capable of being induced in the setting of chronic persistent viral infection.

Our studies in a single individual indicate that virus-specific CD8⁺ T-cell responses that control replication of one strain of HIV-1 may not be sufficiently potent to prevent superinfection with a second virus of the same clade. It must be emphasized, however, that superinfection in an individual already infected with HIV-1 may be different from infection in a person with healthy immune system. In line with studies showing superinfection with a different HIV-1 clade^{23,24}, our results provide a potential explanation for the emergence of a growing number of recombinant viruses worldwide²⁵. In the case described here, although there was only roughly 12% difference in overall sequence, at least half of the targeted CD8 epitopes were not shared between the two viruses, both of which were acquired in North America. Although the immunodominant CD8⁺ T-cell epitope was conserved in both viruses, control of virus B was not achieved until additional epitopes were targeted, which suggests that subdominant responses have a role in immune control. Our study does show the ability of the immune system to generate CD8⁺ T-cell responses against epitopes that have not been previously presented, which indicates that therapeutic immunization in chronic infection may be possible. Further detailed dissection of immune responses in persons who control viraemia and then lose this control should provide important insights regarding the breadth, magnitude and phenotype of HIV-1-specific immune responses that may be necessary to provide broad, cross-protective immunity. □

Methods

Subjects

Individual AC-06 presented with symptomatic acute infection, defined by detectable amounts of plasma HIV-1 RNA (8.8×10^6 copies per ml) and negative antibodies (HIV-1/HIV-2 ELISA and HIV-1 western blot). The CD4⁺ T-cell count at presentation was 551 cells per mm³, and HAART (d4T, 3TC and Nelfinavir) was initiated 4 d later. After 546 d of therapy, he underwent three cycles of STI with the plan to restart therapy if viral load exceeded 5,000 copies of HIV-1 RNA per ml of plasma for three consecutive weeks or 50,000 RNA copies per ml at one time. The fourth STI was off protocol.

IFN- γ Elispot assay

We quantified HIV-1-specific CD8⁺ T-cell responses by Elispot assay as described²¹, using overlapping peptides (peptides with 15–18 residues, overlapping by 10 amino acids) spanning the whole expressed HIV-1 clade-B sequence. Peptides corresponding to optimal clade-B CTL epitopes²⁶ and the autologous virus sequences were also tested. PBMCs were plated at 50,000–100,000 cells per well with peptides at a final concentration of 10^{-5} M in 96-well plates and processed as described²¹. One hundred thousand PBMCs were incubated with media alone (negative control) or phytohaemagglutinin (positive control). The number of specific IFN- γ secreting T-cells was counted by direct visualization, calculated by subtracting the negative control value and expressed as spot-forming cells (SFC) per 10^6 input cells. Negative controls were always less than 30 SFC per 10^6 input cells. A response was considered positive if it was equal to or greater than 50 SFC per 10^6 and at least three times greater than mean background activity. The CD8⁺ T-cell dependence of all responses was confirmed by CD4⁺ T-cell/CD8⁺ T-cell depletion/enrichment studies²¹ using MACS magnetic beads (Miltenyi Biotech). Fine-mapping of epitopes and comparison of recognition of the autologous epitope variants of virus A and virus B were done as described²¹.

HIV-1-specific T_H cell assays

We carried out lymphocyte proliferation assays using baculovirus-derived HIV-1 Gag protein as described³. Briefly, PBMCs were incubated with protein ($5 \mu\text{g ml}^{-1}$) for 6 d and then pulsed with [³H]thymidine at $1.0 \mu\text{Ci}$ per well for 6 h.

Sequencing of autologous virus

Viral DNA was isolated from PBMCs (5×10^6 cells) and viral RNA was extracted from plasma samples and reverse-transcribed as described⁸. A set of seven external primer pairs amplified the whole HIV-1 genome (see Supplementary Fig. 2, Supplementary Table II and ref. 8). Primers were designed on the basis of alignments of over 100 clade-B sequences from the HIV Sequence Database (<http://hiv-web.lanl.gov>). PCR cycling conditions were as follows: 94 °C for 2 min, 35–50 cycles of 30 s at 94 °C, 30 s at 56 °C, 2 min at 72 °C and a final extension of 68 °C for 20 min. RT-PCR cycling conditions were as follows: 50 °C for 60 min, 95 °C for 15 min and cycling conditions as noted above. We used 44 internal primer pairs in nested PCR reactions to yield clonable fragments. The extension times were shortened to 1 min in the nested PCR reactions. PCR fragments were then gel-purified and cloned by TOPO TA (Invitrogen). Plasmid DNA was isolated by 'miniprep' using QiaPrep Turbo Miniprep (Qiagen) and sequenced bidirectionally on an ABI 3100 PRISM automated sequencer (Applied Biosystems). Sequencher (Gene Codes Corp.) and MacVector 4.1 (Oxford Molecular) were used to edit and align sequences. We sequenced a minimum of four clones for each region of the genome at each time point. Neighbour-joining trees were constructed using the Phylip Phylogeny Inference Package Version 3.5c (ref. 27).

Design of specific primers for virus A and virus B

We designed primers specific for virus A and virus B by taking advantage of unique nucleotide substitutions present in each virus in a particular region of Gag (Supplementary Table I). The RT and RT-PCR conditions using these specific primers were the same as those used for the clade-B consensus primers.

Neutralization assays

We measured neutralizing antibodies in human plasma as described²⁰. Titres are reported as the reciprocal plasma dilution that resulted in an 80% reduction in p24 synthesis. Viruses were isolated by PBMC co-culture as described²⁰ and used as original co-culture supernatants in the neutralization assay. All plasma samples were heat-inactivated at 56 °C for 1 h before use.

Statistical analysis

Statistical analysis and graphical presentation was done using SigmaPlot 5.0 (SPSS Inc.). The reported significance (*P* values) of the results is based on two-tailed *t*-tests.

Received 5 July; accepted 13 September 2002; doi:10.1038/nature01200.

- Lisiewicz, J. *et al.* Control of HIV despite the discontinuation of antiretroviral therapy. *N. Engl. J. Med.* **340**, 1683–1684 (1999).
- Ortiz, G. M. *et al.* HIV-1-specific immune responses in subjects who temporarily contain virus replication after discontinuation of highly active antiretroviral therapy. *J. Clin. Invest.* **104**, R13–R18 (1999).
- Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**, 523–526 (2000).
- Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. Virus-specific CD8⁺ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* **68**, 6103–6110 (1994).
- Jin, X. *et al.* Dramatic rise in plasma viremia after CD8⁺ T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189**, 991–998 (1999).
- Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68**, 4650–4655 (1994).
- Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by CD8⁺ lymphocytes. *Science* **283**, 857–860 (1999).
- Allen, T. M. *et al.* Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* **407**, 386–390 (2000).
- Barouch, D. H. *et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290**, 486–492 (2000).
- Amara, R. R. *et al.* Control of a mucosal challenge and prevention of AIDS by a multiprotein DNA/MVA vaccine. *Science* **292**, 69–74 (2001).
- Shiver, J. W. *et al.* Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. *Nature* **415**, 331–335 (2002).
- Hanke, T. *et al.* Effective induction of simian immunodeficiency virus-specific cytotoxic T lymphocytes in macaques by using a multipitope gene and DNA prime-modified vaccinia virus Ankara boost vaccination regimen. *J. Virol.* **73**, 7524–7532 (1999).
- Allen, T. M. *et al.* Tat-vaccinated macaques do not control simian immunodeficiency virus SIVmac239 replication. *J. Virol.* **76**, 4108–4112 (2002).
- Feinberg, M. B. & Moore, J. P. AIDS vaccine models: challenging challenge viruses. *Nature Med.* **8**, 207–210 (2002).
- Malhotra, U. *et al.* Effect of combination antiretroviral therapy on T-cell immunity in acute human immunodeficiency virus type 1 infection. *J. Infect. Dis.* **181**, 121–131 (2000).
- Oxenius, A. *et al.* Early highly active antiretroviral therapy for acute HIV-1 infection preserves immune function of CD8⁺ and CD4⁺ T lymphocytes. *Proc. Natl Acad. Sci. USA* **97**, 3382–3387 (2000).
- Douek, D. C. *et al.* HIV preferentially infects HIV-specific CD4⁺ T cells. *Nature* **417**, 95–98 (2002).
- Barouch, D. H. *et al.* Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. *Nature* **415**, 335–339 (2002).
- Yu, X. G. *et al.* Consistent patterns in the development and immunodominance of human immunodeficiency virus type 1 (HIV-1)-specific CD8⁺ T-cell responses following acute HIV-1 infection. *J. Virol.* **76**, 8690–8701 (2002).
- Montefiori, D. C., Hill, T. S., Vo, H. T., Walker, B. D. & Rosenberg, E. S. Neutralizing antibodies associated with viremia control in a subset of individuals after treatment of acute human immunodeficiency virus type 1 infection. *J. Virol.* **75**, 10200–10207 (2001).
- Altfield, M. A. *et al.* Identification of dominant optimal HLA-B60- and HLA-B61-restricted cytotoxic

T-lymphocyte (CTL) epitopes: rapid characterization of CTL responses by enzyme-linked immunospot assay. *J. Virol.* **74**, 8541–8549 (2000).

- Del Val, M., Schlicht, H. J., Ruppert, T., Reddehase, M. J. & Koszinowski, U. H. Efficient processing of an antigenic sequence for presentation by MHC class I molecules depends on its neighbouring residues in the protein. *Cell* **66**, 1145–1153 (1991).
- Ramos, A. *et al.* Intersubtype human immunodeficiency virus type 1 superinfection following seroconversion to primary infection in two injection drug users. *J. Virol.* **76**, 7444–7452 (2002).
- Jost, S. *et al.* A patient with HIV-1 superinfection. *N. Engl. J. Med.* **347**, 731–736 (2002).
- McCutchan, F. E. Understanding the genetic diversity of HIV-1. *Aids* **14**, S31–S44 (2000).
- Brander, C., Goulder, P. J. R., *et al.* *HIV Molecular Database* (ed. Korber, B. T. M.) 11–120 (Los Alamos National Laboratory, Los Alamos, NM, 1999).
- Felsenstein, J. *PHYMLIP—Phylogeny Inference Package Version 3.5c*, (Dept. Genetics, Univ. Washington, Seattle, 1989).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank the Advanced Computing Laboratory at the Los Alamos National Lab for providing time on the Nirvana supercomputer for the maximum likelihood phylogenetic analysis, and Tammy Bhattacharya for his advice. This study was supported by the Doris Duke Charitable Foundation (M.A., E.S.R., B.D.W.), the NIH (M.A., E.S.R., B.D.W.), the Foundation for AIDS & Immune Research (M.A.), and the Partners/Fenway/Shattuck Center for AIDS Research (T.A., X.G.Y.). B.D.W. is the recipient of a Doris Duke Distinguished Clinical Scientist Award; B.T.K. and P.J.R.G. are recipients of the Elizabeth Glaser Scientist Award.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.D.W. (e-mail: bwalker@partners.org).

TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*

Andreas Hochheimer, Sharleen Zhou, Shuang Zheng, Michael C. Holmes* & Robert Tjian

Department of Molecular and Cell Biology, Howard Hughes Medical Institute, 401 Barker Hall, University of California, Berkeley, California 94720-3204, USA

Drosophila TATA-box-binding protein (TBP)-related factor 2 (TRF2) is a member of a family of TBP-related factors present in metazoan organisms. Recent evidence suggests that TRF2s are required for proper embryonic development and differentiation^{1–5}. However, true target promoters and the mechanisms by which TRF2 operates to control transcription remain elusive. Here we report the antibody affinity purification of a *Drosophila* TRF2-containing complex that contains components of the nucleosome remodelling factor (NURF) chromatin remodelling complex as well as the DNA replication-related element (DRE)-binding factor DREF. This latter finding led us to potential target genes containing TRF2-responsive promoters. We have used a combination of *in vitro* and *in vivo* assays to show that the DREF-containing TRF2 complex directs core promoter recognition of the proliferating cell nuclear antigen (PCNA) gene. We also identified additional TRF2-responsive target genes involved in DNA replication and cell proliferation. These data suggest that TRF2 functions as a core promoter-selectivity factor responsible for coordinating transcription of a subset of genes in *Drosophila*.

Metazoan organisms have evolved diverse mechanisms to control

* Present address: Sangamo BioSciences, Inc., Point Richmond Tech Center, 501 Canal Blvd., Richmond, California 94804, USA.

the spatial and temporal patterns of gene expression during growth, differentiation and development. It has become increasingly evident that cell-type-specific components of the general transcriptional apparatus, for example the mammalian TFIID component TAF_{II}105 (ref. 6) or the *Drosophila* TAF_{II}80 homologue Cannonball⁷ contribute significantly to tissue-specific and gene-selective transcriptional regulation in metazoan organisms. Recent studies have also established that TBP-related factors like *Drosophila* TRF1 can direct transcription from an alternative core promoter⁸ and a TRF1:BRF complex is required for RNA polymerase III transcription of transfer RNA genes⁹. TRF2 is a third member of the TBP family in *Drosophila*¹⁰ and, like TBP and TRF1, TRF2 was shown to interact with the basal transcription factors TFIIA and TFIIB¹⁰. However, the primary amino acid sequence of the putative TRF2 DNA-binding domain has diverged from TBP and TRF1 and, not surprisingly, TRF2 fails to bind to DNA containing canonical TATA boxes¹⁰. But TRF2 is associated with loci on *Drosophila* chromosomes that are distinct from TBP and TRF1 (ref. 10). This suggested that TRF2 may direct promoter specificity and perhaps coordinate a subset of target genes. Size exclusion chromatography indicated that *Drosophila* TRF2 is likely to be part of a macromolecular complex. Unlike TRF1, *Drosophila* TRF2 has amino-terminal and carboxy-terminal extensions flanking the putative DNA-binding core domain. This suggested that TRF2 may be associated with a set of proteins that are distinct from TBP- and TRF1-associated factors^{9,11,12}. We reasoned that the purification and identification of TRF2-associated factors might enable us to identify TRF2-specific promoters and reveal how TRF2 operates to execute transcriptional specificity.

In the absence of a functional assay allowing conventional purification of TRF2-associated factors we generated a panel of monoclonal antibodies directed against several domains of TRF2 to affinity purify TRF2 and its putative associated subunits. We screened approximately 3,500 hybridomas and isolated a clone that efficiently immunoprecipitated TRF2 and its associated factors from *Drosophila* embryo nuclear extract. A Sarkosyl-eluted complex containing TRF2 was analysed by SDS-polyacrylamide gradient gel electrophoresis (PAGE) which revealed a set of 18 polypeptides with relative molecular mass (*M_r*) ranging from 300K to 29K that co-immunoprecipitated consistently with TRF2 even under very stringent conditions (Fig. 1a).

To identify these TRF2-associated polypeptides we obtained several peptide sequences from each subunit and identified 12 distinct TRF2-associated factors (Supplementary Information). Comparative sequence analysis revealed that a number of these subunits are either transcription factors involved in promoter selectivity or are components of other complexes known to be involved in chromatin remodelling. We also obtained four full-length complementary DNA sequences and two coding sequences as predicted by the Berkeley *Drosophila* Genome Project encoding six additional subunits with unknown function (Supplementary Information).

We noticed that the 80K protein associated with TRF2 is identical to DREF¹³. DREF and its corresponding response element DRE have been well documented to be important for the regulation of cell-cycle and cell-proliferation genes in *Drosophila* (that is, genes for PCNA and the 180K and 73K subunits of DNA polymerase¹⁴). The identification of the promoter-selective DNA-binding protein DREF was intriguing because *Drosophila* TRF2 thus far failed to bind to canonical TATA-box elements, which suggests that TRF2 may cooperate with DREF to execute promoter specificity and perhaps operate like a metazoan sigma factor.

We also found that the 140K protein associated with TRF2 is identical to *Drosophila* ISWI, which is the catalytic ATPase subunit of NURF¹⁵, ACF¹⁶ and CHRAC¹⁷ chromatin remodelling complexes. Moreover, the 55K and 38K proteins associated with TRF2 turned out to be NURF-55/CAF-1 (ref. 18) and NURF-38/inorganic pyrophosphatase¹⁹, respectively. Notably, the peptide sequences

obtained from the three largest (300K, 250K and 230K) proteins associated with TRF2 did not match NURF-301 (ref. 20), suggesting that the presence of some NURF subunits is not merely a result of contaminating NURF in the TRF2 complex. However, analysis of the cDNAs encoding the 190K and 160K proteins associated with TRF2 revealed that both proteins contain conserved sequence motifs for 11 and 5 zinc finger motifs (C₂H₂), respectively (Supplementary Information); these smaller proteins thus resemble factors like the CCCTC-binding factor CTCF, which has been implicated in mediating chromatin-dependent processes such as the regulation of insulator function²¹. It is therefore possible that the TRF2 complex encompasses both promoter-selectivity functions and NURF-like components as well as other activities with distinct subunits and specificity.

Sequence analysis of the cDNA encoding the 65K protein revealed a significant similarity to the RNA-binding protein Rap55 (ref. 22) isolated from *Pleurodeles waltl* and *Xenopus laevis*, whereas sequence analysis of the 70K, 116K and 118K proteins showed no significant similarity to known proteins in the databases. The functional relevance of β -tubulin in the TRF2 complex is at present unclear and the 47K and the 29K polypeptides associated with TRF2 are yet to be characterized.

As an independent verification that TRF2 and DREF are components of a multisubunit complex, we performed co-immunoprecipitation experiments using a monoclonal antibody directed against DREF¹³. Sarkosyl-elution revealed a set of polypeptides that is very similar to the TRF2 complex immunopurified by the anti-TRF2 monoclonal antibody (Fig. 1a). Immunoblotting analysis confirmed the presence of ISWI, DREF, TRF2 and NURF-55 in both co-immunoprecipitations (Fig. 1b). Although the stoichiometry of some of the other subunits appear variable in the two independent co-immunoprecipitations, all the putative subunits of the TRF2 complex except for NURF-38 and the 29K polypeptide were consistently present even under the most stringent washing conditions.

Having identified DREF as a tightly associated component of the TRF2 complex, we next asked whether TRF2 can function as a true core promoter recognition factor and selectively initiate tran-

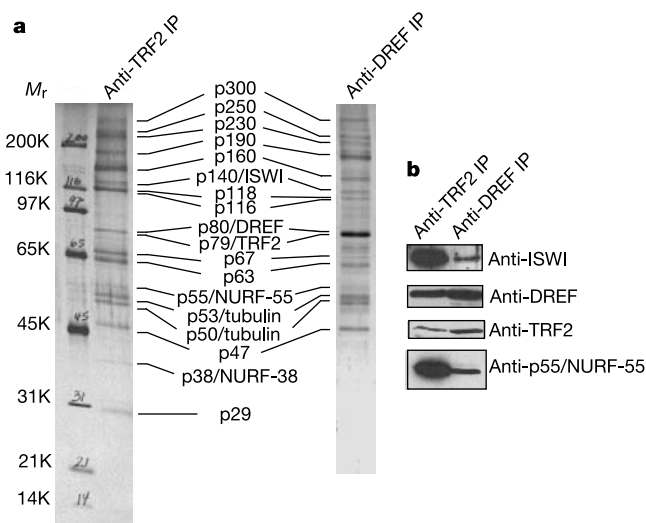


Figure 1 Purification of a *Drosophila* TRF2 complex by co-immunoprecipitation. polyacrylamide gradient gel electrophoresis (PAGE) **a**, Co-immunoprecipitation using anti-TRF2 and anti-DREF monoclonal antibodies. Polypeptides were eluted from the antibody-beads with Sarkosyl, resolved by 7–15% SDS–PAGE and visualized by silver staining. **b**, Immunoblotting analysis of the TRF2 complex using antisera raised against ISWI, DREF, TRF2 and NURF-55. IP, immunoprecipitation.

scription at a promoter that is documented to be stimulated by the DRE/DREF system. We chose the DREF-responsive *PCNA* promoter¹³ which contains at least three promoter-proximal regulatory elements including an upstream regulatory element URE, DRE and two E2F recognition sites located within 200 bp upstream of the start site²³.

To test the responsiveness of the *PCNA* promoter *in vitro* and to map the transcription start site(s) we tested a -580 *PCNA* (-580 to $+56$) promoter fragment, which contains all known regulatory elements, and a -64 *PCNA* (-64 to $+56$) promoter fragment, which lacks all regulatory elements except for the E2F-binding sites as DNA templates for *in vitro* transcription. We added increasing amounts of a partially purified *Drosophila* embryo nuclear extract (H.4) that contains all the necessary basal factors as well as both the TRF2 complex and limiting amounts of TFIID to the transcription reaction (Fig. 2a). Using the -580 *PCNA* template two distinct transcription start sites separated by 63 nucleotides were detected. Promoter 1 (with start site at position $+1$) was stimulated with increasing amounts of H.4 supplemented with TFIID whereas promoter 2 (with start site at position -63) was detected only with the lowest amounts of H.4 added (Fig. 2a, lanes 1–4). Using the truncated -64 *PCNA*, template transcription from promoter 2 was essentially abolished, whereas a weak activity could be detected from promoter 1 by adding the maximum amount of H.4 + TFIID (Fig. 2a, lane 5).

To confirm that both promoter 1 and promoter 2 are functionally

active *in vivo*, we performed primer extension analysis of messenger RNA isolated from SL2 cells stably expressing an inducible TRF2 (Fig. 2b). Transcription initiation from start sites 1 and 2 was observed *in vivo*, each mapping to within 1–2 nucleotides of the *in vitro* sites. Moreover, promoter 2 is stimulated upon overexpression of TRF2 (Fig. 2b, lanes 2 and 3) compared to the uninduced control (Fig. 2b, lane 1) while transcription from promoter 1 remained unchanged. Taken together, these *in vitro* and *in vivo* results suggest that promoter 2 might be TRF2-dependent, whereas promoter 1 appears to be mediated by TFIID.

To substantiate this interpretation we immuno-depleted the H.4 fraction of TFIID then tested the depleted extract for transcription. The TFIID-depleted extract was inactive when using a TFIID-dependent control promoter (*Adh* distal) (Fig. 2c). However, adding increasing amounts of the TFIID-depleted extract to the -580 *PCNA* template selectively potentiated transcription from promoter 2 (Fig. 2a, lanes 6–9) but not from promoter 1. By contrast, no transcription from either start site was observed when using the -64 *PCNA* template (Fig. 2a, lane 10). These results provide further evidence that promoter 2 is likely to be recognized and potentiated by the TRF2 complex in a DRE-dependent manner whereas promoter 1 appears to use the ubiquitous TFIID.

To obtain more direct evidence that TRF2 selectively induces promoter 2 we added increasing amounts of purified recombinant TRF2 and DREF to transcription reactions programmed with the -580 *PCNA* template and the TFIID-depleted extract. Adding

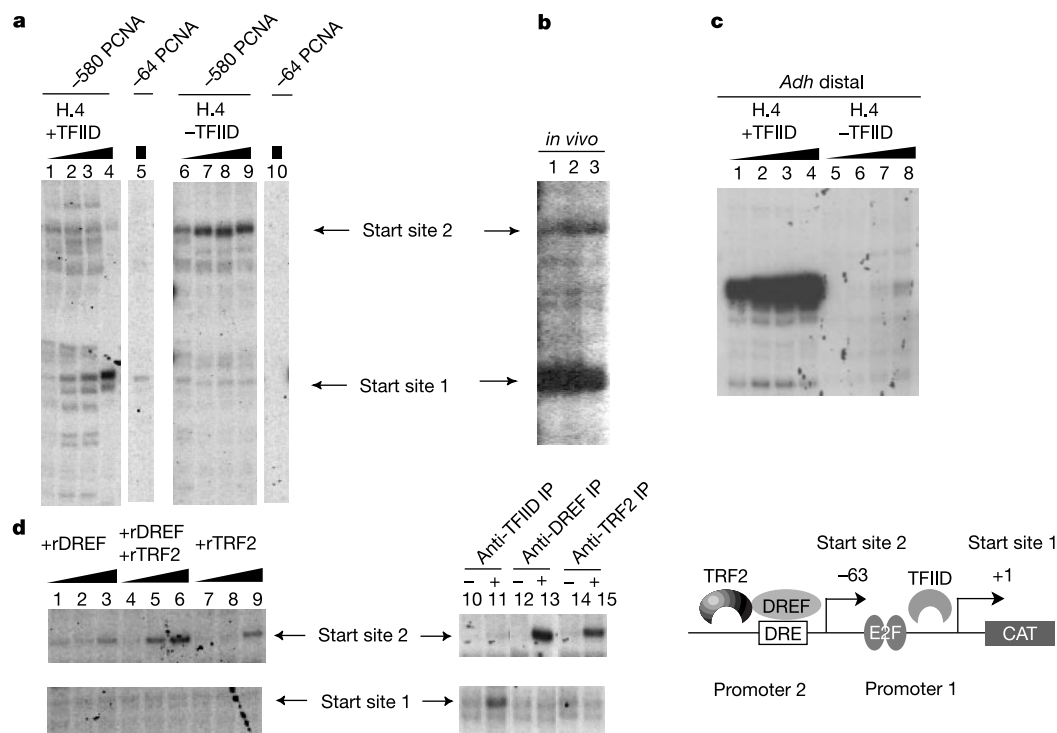


Figure 2 The *PCNA* tandem promoters use two distinct transcription start sites designated start site 1 ($+1$) and start site 2 (-63). **a**, Adding H.4 + TFIID (0.1, 0.25, 0.5 or 0.75 μ l, lanes 1–4) to reconstitute the -580 *PCNA* template *in vitro* stimulates start site 1, whereas transcription from start site 2 is reduced. TFIID-depleted H.4-TFIID (0.1, 0.25, 0.5 or 0.75 μ l, lanes 6–9) stimulates only start site 2. The -64 *PCNA* template (no DRE) is almost inactive using H.4 + TFIID (0.75 μ l, lane 5) or H.4 – TFIID (0.75 μ l, lane 10). **b**, *In vivo* start site mapping of *PCNA*. Start site 2 is stimulated by overexpression of TRF2 (lanes 2 and 3) compared to the uninduced control (lane 1), start site 1 is unchanged. **c**, TFIID-depleted H.4 (H.4 – TFIID) (0.1, 0.25, 0.5 or 0.75 μ l, lanes 5–8) is inactive on

the *adh* distal template compared to H.4 + TFIID (0.1, 0.25, 0.5 or 0.75 μ l, lanes 1–4). **d**, Addition of recombinant DREF (rDREF) + rTRF2 (1, 5 or 10 ng each, lane 4–6) to the -580 *PCNA* template programmed with 0.25 μ l H.4 – TFIID efficiently stimulated start site 2 compared to rDREF (1, 5 or 10 ng, lanes 1–3) or rTRF2 (1, 5 or 10 ng, lanes 7–9) alone. Beads with immunopurified TRF2 complex stimulated start site 2 (0.5 μ l beads, immunopurified with anti-TRF2 monoclonal antibody or anti-DREF monoclonal antibody, lanes 12–15), TFIID on beads stimulated start site 1 (0.5 μ l beads, immunopurified with anti-TAF₂₅₀ monoclonal antibody, lanes 10, 11).

TRF2 or DREF individually to the transcription extract was sufficient to partially stimulate promoter 2 (Fig. 2d, lanes 1–3 and 7–9) without much effect on promoter 1. However, adding both TRF2 and DREF stimulated promoter 2 more efficiently (Fig. 2d, lanes 4–6). Adding the complete endogenous TRF2 complex (immunopurified using either the anti-TRF2 or the anti-DREF monoclonal antibody) directly to the transcription reactions again revealed that only promoter 2 was activated (Fig. 2d, lanes 12–15). In contrast, when TFIID immobilized on beads was added back to the transcription reaction, only promoter 1 was stimulated (Fig. 2d, lanes 10, 11). These biochemical complementation experiments indicate that promoter 1 is TFIID-responsive whereas promoter 2 is selectively activated by the TRF2/DREF complex. These findings support the notion that the TRF2 complex can function as a core promoter-selective transcription factor and further suggest that TRF2 and DREF might work together to target a subset of genes containing DREs.

To determine the promoter-selective properties of TRF2 in the context of living cells, we tested whether the TRF2/DREF complex can stimulate the tandem *PCNA* promoters by using luciferase

reporter assays. To better assess the role of the multiple *cis*-control elements of the *PCNA* promoters we fused three fragments (–580 to +56, –120 to +56 and –64 to +56) harbouring different combinations of these regulatory elements to a reporter (Fig. 3b) and assayed luciferase activity in stably transfected SL2 cells. Induction of either TRF2 or DREF resulted in a strong stimulation (15-fold or 7-fold, respectively) of the reporter gene expression in a dose-dependent manner (Fig. 3a, c). As expected, both the TRF2- and DREF-dependent stimulation were abolished when the DRE was deleted (–64 *PCNA*). The activity of the –580 *PCNA* reporter was consistently higher than the –120 *PCNA* reporter when stimulated by TRF2, suggesting that the URE may contribute to the overall promoter activity. As we expected, control SV40 and herpes simplex virus thymidine kinase (HSV TK) promoters failed to respond to either TRF2 or DREF. Expressing the TBP subunit of TFIID resulted in a weak stimulation of the *PCNA* reporter. As we anticipated, the TBP-dependent HSV TK and SV40 promoters were efficiently stimulated by overexpressing TBP (Fig. 3c). Inducing the co-expression of TRF2 and DREF resulted in a strong promoter-selective stimulation of *PCNA* that was dependent on the presence of the DRE (Fig. 3a, c). We speculate that TRF2 could be limiting in SL2 cells relative to DREF and thus overexpression of TRF2 was sufficient to stimulate transcription from the *PCNA* promoters; alternatively, additional components of the TRF2 complex might be required to mediate a synergistic activation by TRF2 and DREF.

We next asked whether TRF2 can contribute to the enhancer-dependent activated transcription of the *PCNA* promoters by E2F and DP, which cooperate in DNA-binding and transcriptional activation²⁴. The co-expression of just the transcriptional activators E2F and DP in the absence of exogenous TRF2/DREF resulted in a substantial transcriptional activation of the *PCNA* reporter (Fig. 3a, c). It is likely that this activation by E2F/DP is mediated by endogenous TRF2, which was detected by western blot analysis (Fig. 3a). We note that this activation was abolished with the –64 *PCNA* reporter, which lacks the DRE-binding sites but still contains the E2F-binding sites. As we expected, inducing the co-expression of all three promoter recognition factors—TRF2, DREF and E2F—resulted in a strong synergistic activation of the *PCNA* promoters (80-fold) (Fig. 3a, c) in a DRE-dependent fashion. These results suggest that in SL2 cells TRF2 and DREF can work together to

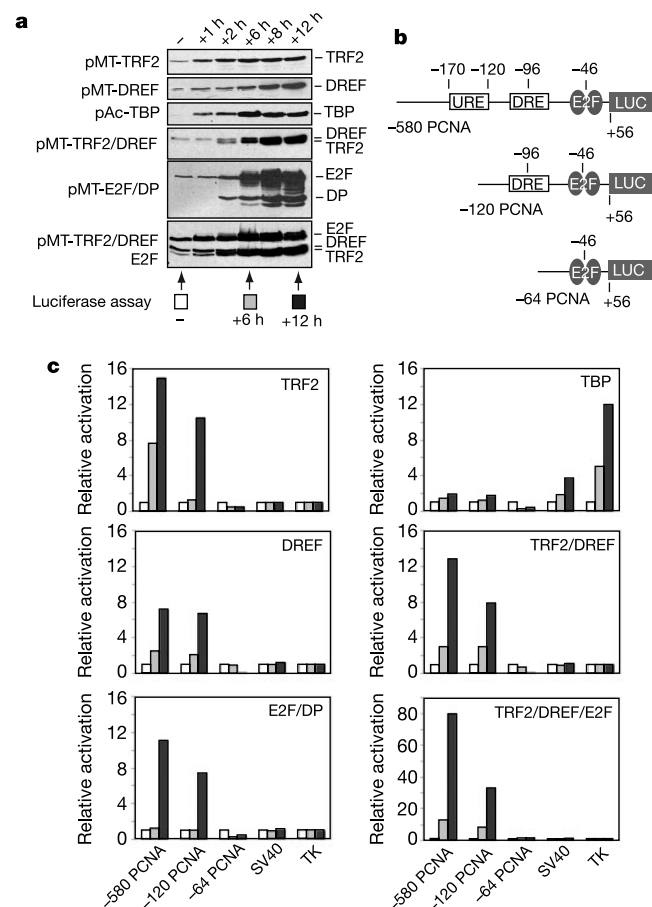


Figure 3 The *PCNA* tandem promoters are synergistically activated by TRF2, DREF and E2F in SL2 cells. **a**, Protein expression levels in stable SL2 cells were monitored by western blot at the indicated time points after induction of protein expression. **b**, Three fragments of the *PCNA* promoter region harbouring different *cis*-regulatory elements were fused to a luciferase reporter. **c**, Luciferase reporter assays show a strong and dose-dependent stimulation of the *PCNA* reporter by TRF2 (15-fold) or DREF (7-fold) in a DRE-dependent manner but not by TBP. The HSV TK and SV40 control promoters are strongly stimulated by TBP but not by TRF2 or DREF. Co-expression of TRF2, DREF and E2F synergistically activates the *PCNA* reporters (80-fold) in a DRE-dependent manner but not the control promoters compared to overexpression of E2F/DP (11-fold) or TRF2/DREF (13-fold). Standard deviation was less than 5%.

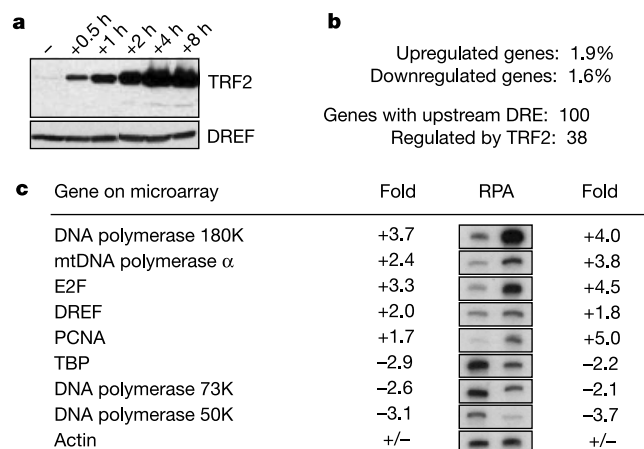


Figure 4 Microarray analysis of TRF2-responsive target genes. **a**, Expression level of TRF2 and DREF was monitored by western blot at time points indicated. **b**, 1.9% of the analysed genes were upregulated and 1.6% downregulated by more than 2-fold upon overexpression of TRF2. *Drosophila* genome analysis revealed 100 genes with a promoter-proximal consensus DRE, 38 of the 100 genes are responsive to TRF2 overexpression. **c**, Microarray analysis revealed TRF2 responsive target genes with a reported *cis*-regulatory DRE confirmed by RNase protection assays (RPA).

stimulate the *PCNA* reporter in a DRE-dependent fashion. This is consistent with the finding *in vitro* that the TRF2 complex can selectively initiate transcription from promoter 2 of the *PCNA* gene in a DRE-dependent manner.

To investigate whether TRF2 is involved in the coordinate regulation of other DNA replication and cell cycle genes in the *Drosophila* genome we hybridized oligonucleotide-based microarrays representing 13,500 *Drosophila* genes with RNA probes isolated at different time points after induction of TRF2 expression in SL2 cells (Fig. 4a). The microarray analysis revealed that only 1.9% of all genes analysed were upregulated more than 2-fold and only 1.6% downregulated by more than 2-fold. Our biochemical studies and cell based assays suggested that TRF2 functions as a core promoter-selectivity factor that collaborates with DREF. We therefore asked whether there are other TRF2-responsive genes that also contain a promoter-proximal DRE. An analysis of the distribution of DREs in the *Drosophila* genome revealed that about 100 genes bear a consensus DRE within 1 kb of the predicted promoter region. Microarray analysis revealed that 38 of these DRE-containing genes were also responsive to TRF2 overexpression (Fig. 4b). For example, genes encoding *PCNA*, the 180K subunit of DNA polymerase, the α -subunit of mitochondrial DNA polymerase, and E2F were all found to be upregulated 2–5-fold by TRF2 in our microarray analysis and confirmed by RNase protection assays. Three additional DRE-regulated genes encoding TBP, the 73K subunit of DNA polymerase, and the 50K subunit of DNA polymerase were found to be downregulated (Fig. 4c).

These data suggest that in addition to the *PCNA* gene a number of other *Drosophila* DRE-containing genes may also be regulated by TRF2 and further support our model that TRF2 can function as a core promoter-selectivity factor that governs a restricted subset of genes that are coordinately regulated. A recent bioinformatics study of core promoter sequences in the *Drosophila* genome identified a consensus DRE as the second most frequent control core element

(other than TATA and INR) providing independent evidence for DRE as a core promoter element (G. Rubin, personal communication).

Because our studies have thus far relied largely on 'gain of function' assays we also employed double stranded RNA interference (dsRNAi) in flies and SL2 cells²⁵ to determine the consequences of ablating TRF2/DREF on transcription of putative target genes such as *PCNA* and *DNApol 180* in cultured cells. We treated SL2 cells with *in vitro* synthesized dsRNA and monitored the depletion of TRF2 and DREF proteins by immunoblot analysis (Fig. 5a). After 2 days of incubation the specifically targeted TRF2 and DREF proteins were severely depleted (Fig. 5a). Mock treatment or a double-stranded control RNA unrelated to TRF2 and DREF had no effect on the target gene products or control proteins (Fig. 5a). We observed that after 48 h of dsRNA treatment, a significant number of cells sloughed off the plate and died (data not shown). This is in accordance with our previous finding that TRF2 RNAi in *Drosophila* embryos is lethal for embryos (Z. Wang and R.T., unpublished data).

However, between 24 and 48 h of dsRNAi treatment we were able to reproducibly measure the activity of transiently transfected luciferase reporters fused to either the *PCNA* or *DNApol 180* promoters and compare them to an internal control reporter gene. The activities from both the *PCNA* (–580 to +56) and *DNApol 180* (–620 to +20) promoters were significantly reduced in TRF2-depleted cells (4.2-fold and 3-fold, respectively) relative to a Renilla luciferase control reporter driven by the *HSV TK* promoter (Fig. 5b). Likewise, these two target gene promoters were downregulated in DREF-depleted cells (3.5-fold and 5-fold, respectively) (Fig. 5b). RNAi treatment using an unrelated double-stranded control RNA had no effect on the activity of either the *PCNA* or *DNApol 180* promoters (Fig. 5b). These depletion experiments using dsRNAi thus support our previous findings that TRF2 and DREF participate in directing transcription of a select subset of genes that include *PCNA* and *DNApol 180*.

Although the dsRNAi studies provide an independent line of evidence to support the notion that TRF2/DREF play a role in promoter selectivity *in vivo*, they fail to provide a direct mechanistic link between TRF2/DREF and the *PCNA* and *DNApol 180* promoters. We therefore carried out chromatin immunoprecipitation (ChIP) experiments to determine the occupancy of TRF2/DREF at the *PCNA* and *DNApol 180* promoters in formaldehyde-treated SL2 cells using antibodies raised against TRF2, DREF and TBP. The precipitated DNA fragments with an average length of 500–1,000 base pairs (bp) were analysed directly by polymerase chain reaction (PCR) (Fig. 5c). The *PCNA* promoter region was specifically precipitated by anti-TRF2 and anti-DREF and to a lesser extent by anti-TBP (Fig. 5c), which is consistent with our previous findings and confirms that TRF2, DREF and the TFIID subunit TBP can co-localize and directly interact with the *PCNA* promoter region in living cells. As further evidence for the targeting of specific promoters by TRF2 and DREF we also analysed the DRE-containing *DNApol 180* promoter region which is selectively precipitated by the TRF2-, DREF- and TBP-specific antibodies (Fig. 5c). As a specificity control we analysed the *actin* promoter region which is precipitated only by the TBP-specific antibody but not by anti-TRF2 and anti-DREF (Fig. 5c). These ChIP experiments strongly support our previous *in vitro* transcription results as well as the cell-based dsRNAi transcription assays in establishing a core promoter selectivity for the TRF2/DREF complex.

In this study we have described the purification of a TBP-related factor 2 (TRF2)-containing complex and the identification of various associated subunits. Notably, the DNA replication-related element (DRE)-binding factor DREF was found as a tightly bound component of the TRF2 complex. Both *in vitro* and *in vivo* evidence suggest that *Drosophila* TRF2 works in concert with DREF to initiate transcription at select genes (*PCNA*, *DNApol 180*) by specifically

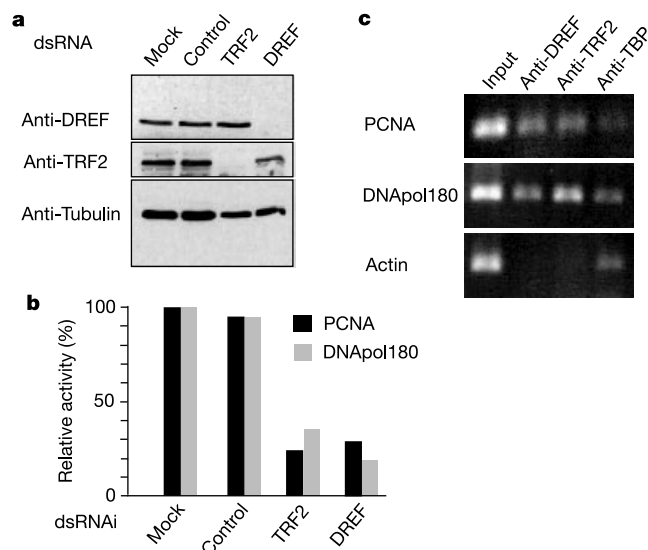


Figure 5 Double-stranded RNA interference in SL2 cells and promoter occupancy of TRF2 and DREF. **a**, dsRNA-mediated depletion of TRF2 and DREF in *Drosophila* SL2 cells was monitored after 48 h by western blot. Control treatment with water (mock) or unrelated double-stranded RNA (control). **b**, Luciferase reporter assays in dsRNA-treated SL2 cells. Activities from both *PCNA* and *DNApol 180* promoters are strongly reduced after dsRNA-mediated depletion of TRF2 (4.2-fold and 3-fold) and DREF (3.5-fold and 5-fold) relative to the *HSV TK* internal control reporter. Standard deviation was less than 8%. **c**, PCR analysis of DNA obtained by chromatin immunoprecipitation of formaldehyde-treated SL2 cells. TRF2, DREF, and the TFIID subunit TBP colocalize to the *PCNA* and *DNApol 180* promoter regions. The *actin* promoter region is occupied by TBP but not by TRF2 and DREF.

targeting a subset of core promoter elements. In the case of the *PCNA* gene, both *in vitro* and cell-based experiments revealed two tandem core promoters (start sites 1 and 2) that are separated by 63 nucleotides. One of the tandemly arranged promoters (site 1) appears to be recognized by TBP/TFIID, whereas the second core element (site 2) is responsive to the TRF2/DREF complex. dsRNAi assays and ChIP experiments further established that TRF2 and DREF co-localize to the *PCNA* and *DNApol 180* promoter regions and that efficient transcription of the two target genes *in vivo* is dependent on the TRF2/DREF complex. The TRF2 complex may therefore serve as a promoter selectivity factor that targets a specific subset of coordinately regulated genes. Indeed, whole-genome microarray analysis revealed that TRF2 selectively regulates the expression of a small subset of genes with promoter elements that contain a DRE including *PCNA* and *DNApol 180*. Many of these putative target genes of TRF2/DREF have been implicated in DNA replication and cell proliferation. These results provide evidence that promoter-selective ensembles like the TRF2 complex may recognize a unique subset of promoters and supports the hypothesis that core promoter diversity combined with specialized components of the basal transcription machinery (that is, TBP-related factors and tissue-specific TAFs) may contribute to a combinatorial mechanism²⁶ that has evolved and diversified to accommodate the elaborate transcriptional networks characteristic of metazoan organisms.

Our studies have thus far not taken into account other functional properties of the TRF2 complex that may potentiate chromatin transactions. Although we have not addressed the biochemical role of putative chromatin remodelling components of the TRF2 complex in this study, we speculate that these components (ISWI, NURF-55, NURF-38 and so on) may contribute additional important functions to efficiently activate transcription of appropriate target genes such as *PCNA* in the context of a chromatin template. Future experiments will be required to determine the biochemical activity and functional specificity of the various subunits in the TRF2 complex. □

Methods

Generation of monoclonal antibodies

Spleen cells of mice with a highly specific anti-TRF2 immune response were fused to myeloma cells essentially as described²⁷. Hybridoma cell clones were screened by enzyme-linked immunosorbent assay (ELISA), western blot analysis and co-immunoprecipitations followed by SDS-PAGE and silver staining.

Immunoprecipitation

Antibody-saturated Protein-G Sepharose (Amersham-Pharmacia) was mixed with nuclear extract derived from 1–12 h *Drosophila* embryos²⁸ in 0.1 M KCl HEMG (25 mM HEPES pH 7.6, 12.5 mM MgCl₂, 0.1 mM EDTA, 10% glycerol, 1 mM dithiothreitol, 1 mM phenylmethyl sulphonyl fluoride (PMSF), and 1 mM sodium metabisulphite) at 4 °C (2 h), washed extensively with 0.15 M KCl HEMG buffer and 0.9 M NaCl HEMG plus 0.1% NP-40, and re-equilibrated to 0.15 M NaCl HEG or 0.15 M KCl HEMG for *in vitro* transcriptions. Complexes were eluted with 0.15 M NaCl HEG plus 0.2% Sarkosyl at 4 °C (1 h) and analysed by 7–15% SDS-PAGE with subsequent silver staining.

Peptide microsequencing

Subunits of immunopurified TRF2 complex were separated by SDS-PAGE and transferred to nitrocellulose. After Ponceau S staining, the bands were excised followed by Lys-C digestion. High-performance liquid chromatography (HPLC) purified peptides were subjected to microsequencing by Edman degradation.

Generation of stable SL2 cell lines and luciferase reporter assays

Genes encoding TRF2, DREF, E2F and DP were cloned into pMT-V5-His (Invitrogen) under the control of the Cu²⁺-inducible *metallothioneine* promoter. SL2 cells (3 × 10⁶) were co-transfected with expression plasmid and pCoHygro (Invitrogen, ratio 20:1) and selected in the presence of 250 µg ml⁻¹ hygromycin B. Clonal lines were screened by immunoblots for high expression and low endogenous background levels. Stable SL2 lines were co-transfected with additional expression plasmids and pCoBlast (Invitrogen, ratio 20:1), selected in the presence of 200 µg ml⁻¹ hygromycin B and 25 µg ml⁻¹ blasticidin, and clonal lines screened for high expression and low endogenous background levels. *PCNA* promoters were fused to the Firefly luciferase in pGL3-Basic (Promega). Cells (3 × 10⁶) were transfected with 200 ng of *PCNA*-luciferase reporter and 50 ng of cytomegalovirus promoter Renilla luciferase control. Luciferase activity was determined after induction of protein expression using DUAL-Luciferase reporter assay system

(Promega) at the time points indicated. Each transfection was performed in triplicate in several independent experiments.

In vitro transcription

PCNA promoters were fused to CAT reporter in pCAT3 (Promega) and purified (CsCl₂ gradient). Nuclear extract derived from 1–12 h *Drosophila* embryos²⁸ was fractionated on a Poros Heparin column. The H.4 fraction was stepped off with 0.4 M KCl HEMG (see Immunoprecipitation section) and dialysed back to 0.1 M KCl HEMG. H.4-TFIID fraction was immuno-depleted using resin-bound anti-TAF_{II}250 monoclonal antibody, and the depletion was monitored by *in vitro* transcription on the *adh* distal promoter template²⁸. The resin was washed with 0.5 M KCl HEMG plus 0.1% NP-40 and re-equilibrated to 0.15 M KCl HEMG. *In vitro* transcription was performed as described²⁸. Reactions were pre-incubated for 10 min (25 °C), initiated with 0.5 mM ribonucleotides and extended for 20 min (25 °C) followed by RNA isolation and primer extension analysis.

In vivo start-site mapping of PCNA

Primer extension analysis was performed with γ-³²P-labelled oligonucleotide PE-a (AGGCGTGCCTCGAACATATTG) hybridized to poly A mRNA derived from stably transfected SL2 cells 0, 4 and 8 h after induction of TRF2 overexpression.

Chromatin immunoprecipitation

SL2 cells (2 × 10⁶ cells ml⁻¹) were treated with 1% formaldehyde (15 min), lysed in lysis buffer (50 mM HEPES pH 7.6, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.5 mM PMSF), followed by ultracentrifugation (30 min, 96,000g). Pellets were resuspended in 1 ml lysis buffer per 2 × 10⁶ cells and sonicated (8 × 30 s) to obtain chromatin fragments (500–1,000 bp). Extracts were spun (15 min, 27,000g) and filtrated (0.22 µm). Precleared extracts (1 × 10⁸ cells) were incubated with 1 µg antibody (2 h) then 1 h with 25 µl protein G-Sepharose. Resin was washed twice with lysis buffer, once with lysis buffer containing 500 mM NaCl, twice with 0.1 M Tris-HCl pH 7.9, 0.25 M LiCl, 0.5% NP-40, 0.5% Na-deoxycholate, 0.5 mM PMSF, and once with 50 mM Tris-HCl pH 7.9, 1 mM EDTA. Chromatin precipitate was eluted, the crosslinks reversed with 250 µl 50 mM Tris-HCl pH 7.9, 10 mM EDTA, 0.8% SDS (65 °C, 12–14 h), and treated with proteinase K. DNA was isolated by phenol-chloroform extraction/ethanol precipitation, followed by RNaseA treatment and phenol-chloroform extraction/ethanol precipitation. Chromatin extract (50 µl) was treated equally and served as input control. Standardized PCR (50 µl) contained 1.0 µl ChIP DNA or input control DNA, 5 pmol oligonucleotide primers, 200 µM NTPs, 1.5 mM MgCl₂, and 2.5 U Taq polymerase. 94 °C for 3 min, then 30 amplification cycles: 94 °C (30 s), 53–58 °C (1 min), 72 °C (40 s) and a final extension at 72 °C (7 min).

Microarray and RNase protection assays

Total RNA was purified using TRI reagent (Sigma) and poly A mRNA purified using Oligotex (Qiagen). Generation of Biotin-labelled cRNA and probing of Affymetrix GeneChips was performed essentially as described²⁹. Hybridizations were normalized by comparing induced culture with the uninduced culture. Sense and antisense riboprobes labelled with ³²P-uridine 5'-triphosphate for RNase protection assays were synthesized with Maxiscript (Ambion). Protection assays were carried out with RPA III (Ambion) using poly A mRNA isolated from stable SL2 cells before and 4 h after induction of TRF2 overexpression.

Double-stranded RNA interference

Sense and antisense RNAs for TRF2 (amino acids 1–194) and DREF (amino acids 1–212) were synthesized from T7 and T3 priming sites with Megascript (Ambion). dsRNAi in *Drosophila* SL2 cells was performed as described²⁵. RNAi-treated cells were transfected with 200 ng of Firefly luciferase reporter and 50 ng of *HSV TK*-Renilla control. Luciferase activity was determined using DUAL-Luciferase reporter assay system (Promega).

Received 8 May; accepted 27 September 2002; doi:10.1038/nature01167.

1. Dantonel, J. C., Quintin, S., Lakatos, L., Labouesse, M. & Tora, L. TBP-like factor is required for embryonic RNA polymerase II transcription in *C. elegans*. *Mol. Cell* **6**, 715–722 (2000).
2. Kaltenbach, L., Horner, M. A., Rothman, J. H. & Mango, S. E. The TBP-like factor CeTLF is required to activate RNA polymerase II transcription during *C. elegans* embryogenesis. *Mol. Cell* **6**, 705–713 (2000).
3. Martjanov, I. et al. Late arrest of spermiogenesis and germ cell apoptosis in mice lacking the TBP-like TLF/TRF2 gene. *Mol. Cell* **7**, 509–515 (2001).
4. Veenstra, G. J., Weeks, D. L. & Wolffe, A. P. Distinct roles for TBP and TBP-like factor in early embryonic gene transcription in *Xenopus*. *Science* **290**, 2312–2315 (2000).
5. Zhang, D., Penttilä, T. L., Morris, P. L., Teichmann, M. & Roeder, R. G. Spermiogenesis deficiency in mice lacking the Trf2 gene. *Science* **292**, 1153–1155 (2001).
6. Freiman, R. N. et al. Requirement of tissue-selective TBP-associated factor TAF_{II}105 in ovarian development. *Science* **293**, 2084–2087 (2001).
7. Hiller, M. A., Lin, T. Y., Wood, C. & Fuller, M. T. Developmental regulation of transcription by a tissue-specific TAF homolog. *Genes Dev.* **15**, 1021–1030 (2001).
8. Holmes, M. C. & Tjian, R. Promoter-selective properties of the TBP-related factor TRF1. *Science* **288**, 867–870 (2000).
9. Takada, S., Lis, J. T., Zhou, S. & Tjian, R. A TRF1: BRF complex directs *Drosophila* RNA polymerase III transcription. *Cell* **101**, 459–469 (2000).
10. Rabenstein, M. D., Zhou, S., Lis, J. T. & Tjian, R. TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. *Proc. Natl Acad. Sci. USA* **96**, 4791–4796 (1999).
11. Goodrich, J. A. & Tjian, R. TBP-TAF complexes: selectivity factors for eukaryotic transcription. *Curr. Opin. Cell Biol.* **6**, 403–409 (1994).
12. Hansen, S. K., Takada, S., Jacobson, R. H., Lis, J. T. & Tjian, R. Transcription properties of a cell type-specific TATA-binding protein, TRE. *Cell* **91**, 71–83 (1997).

13. Hirose, F. *et al.* Isolation and characterization of cDNA for DREF, a promoter-activating factor for *Drosophila* DNA replication-related genes. *J. Biol. Chem.* **271**, 3930–3937 (1996).
14. Hirose, F., Yamaguchi, M., Handa, H., Inomata, Y. & Matsukage, A. Novel 8-base pair sequence (*Drosophila* DNA replication-related element) and specific binding factor involved in the expression of *Drosophila* genes for DNA polymerase alpha and proliferating cell nuclear antigen. *J. Biol. Chem.* **268**, 2092–2099 (1993).
15. Tsukiyama, T., Daniel, C., Tamkun, J. & Wu, C. ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell* **83**, 1021–1026 (1995).
16. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
17. VargaWeisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598–602 (1997).
18. Martínez-Balbás, M. A., Tsukiyama, T., Gdula, D. & Wu, C. *Drosophila* NURF-55, a WD repeat protein involved in histone metabolism. *Proc. Natl Acad. Sci. USA* **95**, 132–137 (1998).
19. Gdula, D. A., Sandaltzopoulos, R., Tsukiyama, T., Ossipow, V. & Wu, C. Inorganic pyrophosphatase is a component of the *Drosophila* nucleosome remodeling factor complex. *Genes Dev.* **12**, 3206–3216 (1998).
20. Xiao, H. *et al.* Dual functions of largest nurf subunit nurf301 in nucleosome sliding and transcription factor interactions. *Mol. Cell* **8**, 531–543 (2001).
21. Bell, A. C., West, A. G. & Felsenfeld, G. Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science* **291**, 447–450 (2001).
22. Lieb, B., Carl, M., Hock, R., Gebauer, D. & Scheer, U. Identification of a novel mRNA-associated protein in oocytes of *Pleurodeles waltl* and *Xenopus laevis*. *Exp. Cell Res.* **245**, 272–281 (1998).
23. Yamaguchi, M., Hirose, F. & Matsukage, A. Roles of multiple promoter elements of the proliferating cell nuclear antigen gene during *Drosophila* development. *Genes Cells* **1**, 47–58 (1996).
24. Dynlacht, B. D., Brook, A., Dembski, M., Yenush, L. & Dyson, N. DNA-binding and trans-activation properties of *Drosophila* E2F and DP proteins. *Proc. Natl Acad. Sci. USA* **91**, 6359–6363 (1994).
25. Clemens, J. C. *et al.* Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl Acad. Sci. USA* **97**, 6499–6503 (2000).
26. Smale, S. T. Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.* **15**, 2503–2508 (2001).
27. Coligan, J. E., Margulies, D. H., Shevach, E. M. & Strober, W. (eds) *Current Protocols in Immunology* (John Wiley & Sons, New York, 1991).
28. Heberlein, U. & Tjian, R. Temporal pattern of alcohol dehydrogenase gene transcription reproduced by *Drosophila* stage-specific embryonic extracts. *Nature* **331**, 410–415 (1988).
29. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Haggart and R. Bell for technical assistance; D. Fyodorov and J. Kadonaga for providing antibodies raised against Acf, ISWI and p53/NURF-55; F. Hirose for generously providing Hybridoma cell lines producing anti-DREF antibody; T. Orr-Weaver for generously providing antibody raised against E2F prior to publication; E. Morris and N. Dyson for providing E2F and DP plasmids; H. Xiao and C. Wu for analysing NURF-301 for peptide sequence matches before publication; and Y. Isogai for providing affinity purified antibodies for ChIPs. We also thank Y. Berghöfer-Hochheimer, R. Freiman, B. Glover, Y. Isogai, T. Kutzkey and B. Lemon for comments on the manuscript, and all members in our laboratory for discussions and technical advice. A.H. was in part supported by a research fellowship by the Deutsche Forschungsgemeinschaft (DFG).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.T. (e-mail: jmlim@uclink4.berkeley.edu).

errata

The prolyl isomerase Pin1 is a regulator of p53 in genotoxic response

Hongwu Zheng, Han You, Xiao Zhen Zhou, Stephen A. Murray, Takafumi Uchida, Gerburg Wult, Ling Gu, Xiaoren Tang, Kun Ping Lu & Zhi-Xiong Jim Xiao

Nature **419**, 849–853 (2002).

The last sentence of the Acknowledgements of this Letter should read “This work was supported by the NIH (Z.-X.(J.)X., K.P.L.) and the Department of Defense (Z.-X.(J.)X.).” □

Patched acts catalytically to suppress the activity of Smoothened

J. Taipale, M. K. Cooper, T. Maiti & P. A. Beachy

Nature **418**, 892–897 (2002).

In the US printing only of this issue of *Nature*, panels c and d of Fig. 2 in this Letter appeared much darker than intended. The immunofluorescence in these panels shows the distinct subcellular localizations of Ptc and Smo in cells, which support the argument against stoichiometric interaction. We recommend that readers view the online version of the paper on *Nature's* website (<http://www.nature.com/nature>). □

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silke Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Hideki Watanabe

e-mail: h.watanabe@naturejp.com

naturejobs

Clearing your own path

Several recent European career forums have concluded that alternative scientific pathways need to be more clearly defined. But Gail Cardew, head of programmes at the Royal Institution in London, believes that scientists who want to pursue a career beyond the bench need to forge their own path. This process requires trial and error, according to Cardew, as well as some self-analysis.

Speaking at an alternative-career forum sponsored by *Naturejobs* at the International Biotechnology Exhibition and Conference in London last week, Cardew explained how she had created her career path. At graduate school, she realized she didn't want to focus in detail on one research project. So she switched to publishing — editing journal articles and keeping in touch with science by attending conferences.

It was then that Cardew became interested in dialogue rather than the written word — in particular, the way in which scientists deal with the public. She felt that most were following the “condescending model”, which assumes that if only the public understood more about science, then controversy over issues such as genetically modified food would simply fade away. She took her current post in an effort to create a different approach, and she now acts as an interface between the government and the public over scientific issues. Although she says that she “stumbled into” her vocation, she feels that her job now meets all of her criteria.

Graduate students no longer need to stumble quite so much, Cardew says. For example, UK students considering a similar path can join the Association of British Science Writers or the Science, Technology, Engineering and Medicine Public Relations Association to see whether a particular career alternative is a good fit. But they should also create a dialogue with themselves. “People should play to their skills, their strengths and their interests,” Cardew says.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

The rise of the United States provokes a European immunological response p4

Japan takes immunology back to basics p7

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

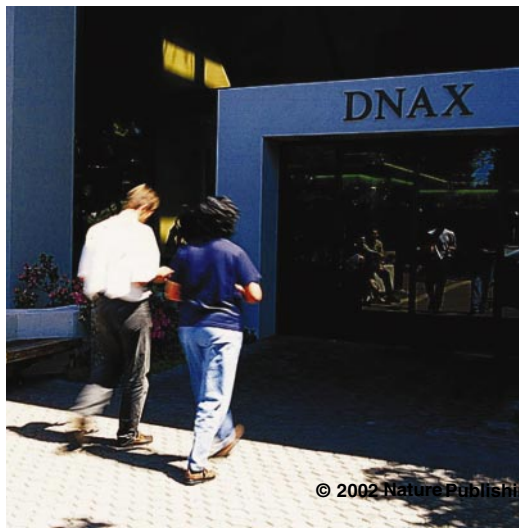
Groundbreaking research

An increasing emphasis on immunology offers fresh opportunities, but they may be tilted towards applied research and the United States, says Virginia Gewin.

The Institute of Immunology currently being built in Marseille-Luminy, France, is one of many such initiatives that are breathing life back into European immunology.

With a US\$1.8-billion biodefence research agenda, more than \$2 billion of Bill & Melinda Gates Foundation money, a host of new technologies and a push to integrate immunology, the United States is flush with opportunities in this field — so much so that it is causing recruitment woes to ripple across the Atlantic. Young immunologists from around the world are taking advantage of the US boom, leaving Europe with a skills shortage. New European Community (EC) programmes and individual initiatives are attempting to stem the losses. But the public sector is following the private sector in a renewed emphasis on applied rather than basic research.

Shift in emphasis: many of DNAX's immunologists sought fresh opportunities when the institute moved away from basic research.



DNAX, formerly a biotech research institute based in San Francisco, is one of several companies started by biotechnology pioneer Alejandro Zaffaroni. Once acquired by Schering-Plough, a New Jersey-based drug company, DNAX became more commercially oriented. Although it still has a basic-research programme, including 30 postdocs, its emphasis has shifted to developing marketable therapeutics — a move that prompted a number of its immunologists to seek other research opportunities.

The DNAX story resonates with other big pharmaceutical trends. Although a faltering economy has slowed hiring at some of the big drug companies, industry still offers roughly 80% of the jobs in immunology. Many think that the economic downturn is temporary and the five-year forecast is good. "This is the decade of immunology," says Glenn Gormley, head of clinical development for drug company AstraZeneca in the United States, "the time that we see an explosion in the practical applications of immunology."

Even in academia, more emphasis is being put on product-driven research. Some are concerned, for example, that despite increased funding for the National Institute of Allergy and Infectious Diseases (NIAID), the US government's biodefence directives may require more money to be spent on development than on research (see *Naturejobs* 4–5; 18 April 2002).

But many researchers are confident that the NIAID has outlined a portfolio that will balance basic-research needs with immediate interests. The biodefence funds cover some exciting areas, such as

DNAX

Not immune to alternatives

Alternative career tracks are a big topic among students these days.

"Recently, when I give talks, an underwhelming number of students actually expect to go into academia," says Michele Hogan, president of the American Association of Immunologists.

Given that most academic positions become available only when someone retires, competition is tough. Paul Kincade, head of the immunobiology and cancer research programme at the Oklahoma Medical Research Foundation, advocates a restructuring of opportunities in the field. He suggests that institutions should invest more to develop stable, long-term

contracts, making non-tenure-track staff scientist positions a more attractive alternative.

Other students are anticipating the problem, and avoiding academia. "I have a few who have gone into law school or venture capitals," says Tak Mak, an immunologist at the University of Toronto in Canada.

Increasingly, immunologists are keeping one foot in business and the other in academia. Pramod Srivastava, an expert in heat-shock proteins and immune responses, founded the New York drug-discovery company Antigenics but is also professor of immunology at the University of Connecticut School of Medicine.

V.G.

picture' immunologists are hard to find. "If you hire someone because of a skill and not a demonstrated thought process, they are eventually not going to need that skill," he says.

Similar complaints are expressed in academia. "I consistently hear that it is hard to get good postdocs," says Michele Hogan, executive director of the American Association of Immunologists. It is an applicant's market with plenty of work available and plum jobs on offer to immunologists with the right breadth of knowledge and creativity.



Tak Mak wants to see more basic research in immunology to tackle diseases such as malaria and tuberculosis.

TIGHTENING BELTS

In the United Kingdom, where there is little government funding for immunology, many scientists admit that recruitment is becoming problematic. Anne O'Garra, a former DNAX employee who is now a consultant, has recently returned to the United Kingdom to head the new Laboratory of Immunoregulation at the National Institute for Medical Research in north London. She doesn't paint as bleak a picture as some. Within this new division, she will recruit independent investigators to integrate the fields of molecular immunology, immune-cell biology, infectious disease, parasitology, virology and mycobacterial research over the next few years — proof positive that these fields are merging as knowledge is gained. But she admits that academic research opportunities are more limited.

The lack of government funding leaves many scientists pessimistic, as they rely primarily on the state and the medical charity the Wellcome Trust for cash to conduct their research. "Without the Wellcome Trust, UK biomedical research is not even worth thinking about," says Eddy Liew, head of the immunology department at the University of Glasgow, UK.

Although the Wellcome Trust does not fund basic-science fellowships for first-year postdocs, it does offer advanced training fellowships and research career-development fellowships to senior postdocs. The trust is currently attempting to make these more attractive by boosting the salaries and increasing the stipends of students. Unfortunately, it hasn't been matched by other public bodies. It will take some time before the 'brain drain' trend can be reversed, says Liew.

Having worked in both industry and academia, O'Garra doesn't discount the synergies between the two. Louise Dunn, director of science communications at drug company GlaxoSmithKline, agrees that the interchange of people between public and private efforts is healthy for science overall. But even with heightened activity in those sectors, says Dunn, this could be a lean year for graduates seeking jobs in the pharmaceutical industry because of industry consolidation and a less favourable economic climate than in the recent past.

Basic-research funding in Europe has taken the brunt of the economic hard times. In fact, the abrupt

innate immunity and adjuvant discovery, which are seen by many researchers as ripe for advances in knowledge and could lead to new vaccines.

The philanthropic Bill & Melinda Gates Foundation in Seattle, Washington, provides much-needed funding for research on some of the world's most deadly infectious diseases. "The Gates Foundation tackles problems that require more scientific investigation," says Tak Mak, an immunologist at the University of Toronto, Canada. "We don't have a clue about malaria, HIV or tuberculosis," he adds. Mak anticipates that, with Richard Klausner, former director of the US National Cancer Institute, at the helm, the Gates Foundation will do much to promote strategic research in these areas — and will require a number of good immunology and vaccine people.

BROAD-SPECTRUM APPROACH

To open up these new avenues of applied research — whether in academia or industry — recruiters are looking for immunologists with a broader view. Andrew Chan, senior director of the immunology department at biotechnology company Genentech in South San Francisco, recently concluded 14 months of reviewing more than 1,000 applications in order to fill six critical positions. Although many had the technical expertise, Genentech wants people who will bring "creativity and originality to a scientific programme", says Chan. AstraZeneca's Gormley agrees that 'big-



Jill O'Donnell-Tormey believes a cancer vaccine could be available within 15 years.

Prospects for cancer

The promise of immunology-based cancer therapies is not lost on industry. "I think it is an area still undergoing growth, despite the 15–20 companies currently developing these therapies," says Brian Barber, chief scientific officer and senior vice-president of Mojave

Therapeutics in Hawthorne, New York.

Many believe that success, which will encourage more investment, will come quickly. "An effective vaccine could be in mainstream use within 10–15 years," says Jill O'Donnell-Tormey, executive director of the Cancer Research Institute

(CRI) in New York. To quicken that pace, the CRI recently joined forces with the Ludwig Institute for Cancer Research, also in New York, to create the Cancer Vaccine Collaborative. Realizing that early-stage clinical trials need to be more academic, they have created a programme to standardize results

across multiple trials.

But charities such as the CRI, which provide a substantial number of postdoctoral fellowships, are following a trend started by the US National Institutes of Health to limit the number of senior fellows. "We put preference on funding first-time postdocs," says O'Donnell-Tormey. **V.G.**

dismantling of the Basel Institute for Immunology in Switzerland last year into its new incarnation as the Roche Centre of Medical Genomics signalled the subsequent shift to applied research by US-based companies such as DNAX. Klaus Lindpaintner, director of the Roche centre, says that the move was designed to bring the research closer to issues in drug discovery.

BABY BASELS

The Basel Institute was seen as exceptional because of the autonomy enjoyed by its researchers. Now some scientists, lamenting the loss of the institute and hampered by the lack of government research money, are creating 'baby Basels' elsewhere in Europe. Smaller and more specialized, these research institutes are attempting to provide alternative avenues for scientists early in their careers. Local-government support, which varies from country to country, is crucial for these endeavours.

"In Europe we need to have a plan to promote good young scientists and allow them to develop independent research," says Antonio Lanzavecchia, director of the two-year-old Institute for Research in Biomedicine in Bellinzona, Switzerland. "We're trying to fill the gap left by the Basel Institute."

Although ground has yet to be broken on the Institute of Molecular Biotechnology in Vienna (see *Naturejobs* 4–5; 11 April 2002), the founders have set up shop in a rented laboratory nearby. Josef Penninger, the institute's scientific and administrative director, is actively recruiting for eight principal-investigator positions and support staff. "I believe that 95% of research is done by 5% of people — these are the people we want to hire," says Penninger.

Under Philippe Kourilsky's leadership, France's Pasteur Institute in Paris is also undergoing change. It is now very active in developing new research groups, says James Di Santo, who directs an immunology lab at the institute. To launch up to 50 new research units, the institute has set up a system of recruiting 10 five-year groups per year, two of which are currently devoted to immunology.

At present, getting established in the EC can be a tortuous process. "We have a shortage of positions and a surplus of good people," says Harald Renz, chairman of the immunology division at the European Academy of Allergology and Clinical Immunology (EAACI). "Local

and regional funding opportunities are steadily declining, and, in contrast, European funding possibilities are increasing. It's more and more important that scientists come together on the European level because of this shift in terms of money," says Renz.

In fact, the Sixth Framework Programme — the main instrument for funding scientific research in Europe — emphasizes such collaborations. Although its structure has been criticized, it may be the only lifeline for those EC countries hit hard by decreasing levels of government funding. The programme will be finalized this month, although the preliminary budget allocates over 2.3 billion euros (US\$2.3 billion) for health-related life-science and biotechnology research, including treatments that stimulate the immune system. Competing for that funding will involve forming networks with labs throughout the EC.

Many researchers hope that these networks will jump-start European science. Italian researchers, in particular, are having a hard time securing funding to keep research alive. "Either we will be part of these networks, or there is no survival. It wasn't like this 10 years ago," says Paola Ricciardi-Castagnoli, an immunologist at the University of Milano-Bicocca.

In the next two years, the pool of EC researchers will grow rapidly. Ten eastern-European countries are to join the European Union, bringing with them a stream of students who will compete for the existing positions. "In these countries, there are huge numbers of eager young scientists with good knowledge, but limited access to labs and technology," says Renz. "We have to create some kind of organized exchange programmes to get these people on board." Bernard Malissen, immunologist at the Marseille-Luminy Centre for Immunology in France, looks forward to the new arrivals. "The bigger European science can be, the better it will be," he says.

Art Krieg, chief scientific officer of the Coley Pharmaceutical Group, based in Wellesley, Massachusetts, agrees. "Immunology is such a growth area, there is plenty of room to absorb more scientists," he says. For the time being, it appears that the bulk of that absorption will take place in the United States. ■

Vigina Gewin is a freelance science writer based in Corvallis, Oregon.
 Institute for Research in Biomedicine ♦ www.irb.unisi.ch/index.asp
 Institute of Molecular Biotechnology ♦ www.imba.oeaw.ac.at
 The Wellcome Trust ♦ www.wellcome.ac.uk
 The Bill & Melinda Gates Foundation ♦ www.gatesfoundation.org



Klaus Lindpaintner, director of the Roche Center of Medical Genomics, is focusing on drug discovery.

Back to basics

The new Center for Allergy and Immunology aims to give Japanese immunology a fresh perspective — and better working conditions for young scientists, says Robert Triendl.

Japan is taking immunology back down to single-cell level, as demonstrated by this imaging set-up examining single molecules in a cell.

Japan has a strong history of immunology, although to some degree the country's stolid tradition has also hampered its researchers trying to capitalize on this base. Japanese scientists have been responsible for a number of groundbreaking advances in the field during the past decade. For example, major discoveries in how the protein interleukin-6 recognizes inflammatory insults and marshalls a response originated there in the 1990s. But such successes belie the systemic lack of autonomy, especially for younger scientists; limited support for infrastructure such as animal facilities, cell and tissue repositories, and electronic databases; and a reluctance to invest in and manage large-scale technology-driven projects that involve outside collaborators or contractors.

But that could be about to change. RIKEN, the Institute of Physical and Chemical Research, is currently building the Research Center for Allergy and Immunology (RCAI) at its campus in Yokohama. The strategy behind the new centre is ambitious, says its director, Masaru Taniguchi, a molecular immunologist at Chiba University. The RCAI was founded to alleviate some of the shortcomings of Japanese university-based research in allergy and immunology, and to develop fresh approaches to common immunological problems.

"It is time to go back to a systemic view of immunology and to ask how the immune system is regulated," says Taniguchi. "With the advances in genomics or proteomics we are now in a position to recreate a systemic vision of immunology at the level of a single cell."

Proteomics will aid the discovery of molecules and will help to elucidate the more specific functions of proteins in a cell, says Taniguchi. And the fact that the RCAI will share the Yokohama campus with RIKEN's Genomic Sciences Center can only help it as it pursues its goals. In addition to genomics and proteomics, the RCAI will invest in the development of microarray platforms, bioinformatics, computational biology, new technologies for cell sorting and single-molecule analysis, and a large-scale animal facility.

The new centre will provide young investigators with a chance to concentrate solely on research, without having to deal with the administrative distractions characteristic of university life. But, like the other newly created RIKEN centres, the RCAI will not provide permanent positions — scientists will be hired on five-year, renewable contracts.

The centre has already filled most of its group leader positions. Until next summer, when the Yokohama facilities will be ready, most of the RCAI's scientists will continue to work in their present positions in academia, or move into a laboratory complex rented from Chiba University and filled with equipment for the new centre.

Finding the right recruits to fill these labs was not easy, says Taniguchi. Increased flexibility in Japanese employment would make recruiting top people for temporary posts easier, he says. "We need a rule that allows scientists in tenured position to have joint appointments — and to take a leave of absence to go and work for RIKEN for a couple of years," he adds.

This is not the first time that Taniguchi has dealt with personnel issues. During his time as dean of Chiba University's medical school, he managed to secure a more than threefold increase in the salary for technicians. "You simply don't get capable technicians for the salary of a part-time employee at a convenience store," he points out, adding that nobody had ever tried to challenge the unwritten rules on how to pay technicians. There is little doubt that at RIKEN, his previous management experience will serve him well.

Robert Triendl is a freelance science writer based in Tokyo.

Research Center for Allergy and Immunology

♦ www.rcai.riken.go.jp/indexE.html



Masaru Taniguchi: setting new standards for Japanese immunology.

Once completed, the new Research Center for Allergy and Immunology plans to offer young researchers a distraction-free environment.

